Numpy, Pandas, and Scikit-Learn

1

LearnQuest

Why Use Pandas and Numpy?

1. Already written code for data science (Python libraries)
2. Faster data wrangling and analysis! (Written mostly in C)

Why Use Pandas and Numpy?

1. Already written code for data science (Python libraries)
2. Faster data wrangling and analysis! (Written mostly in C)

Think about writing code to

- **read in from a file**
- **group by specific columns**
- **apply some aggregate function**
- **write to a file**

Why Use Pandas and Numpy?

1. Already written code for data science (Python libraries)
2. Faster data wrangling and analysis! (Written mostly in C)

Think about writing code to

- **read in from a file**
- **group by specific columns**
- **apply some aggregate function**
- **write to a file**

Would be a pain to code from scratch!!!

Why Use Pandas and Numpy?

1. Already written code for data science (Python libraries)
2. Faster data wrangling and analysis! (Written mostly in C)

Think about writing code to

- **read in from a file**
- **group by specific columns**
- **apply some aggregate function**
- **write to a file**

1 line of code with Pandas!

Would be a pain to code from scratch!!!

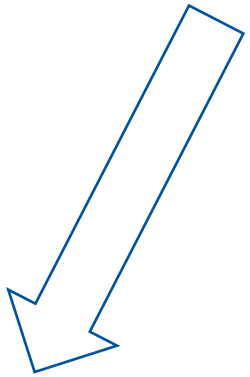
Why Use Pandas and Numpy?

1. Already written code for data science (Python libraries)
2. Faster data wrangling and analysis! (Written mostly in C)

Think about writing code to

- **read in from a file**
- **group by specific columns**
- **apply some aggregate function**
- **write to a file**

1 line of code with Pandas!



Would be a pain to code from scratch!!!

```
pd.read_csv('orders.csv').groupby('Order Date').sum().to_csv('grouped_orders.csv')
```

A Tale of Two Matrices

Numpy

- **Type:** Mathematical Matrices (only numbers)
- **Main structure:** Arrays – similar to Excel with just numbers
- **Use Cases:** Matrix to a 5^{th} power, random numbers, number ranges, missing values, list conversions, linear algebraic calculations
- **Written mostly in:** C

Pandas

- **Type:** Data Science Matrices (numbers and words)
- **Main structure:** Dataframes - similar to SQL Tables
- **Use Cases:** Groupby/sorting, reading/writing files from csv, Excel, Google Drive, SQL, etc.; regression, plotting
- **Written mostly in:** Python and Cython, with critical functions in C

A Tale of Two Matrices

Numpy

- **Type:** Mathematical Matrices (only numbers)
- **Main structure:** Arrays – similar to Excel with just numbers
- **Use Cases:** Matrix to a 5th power, random numbers, number ranges, missing values, list conversions, linear algebraic calculations
- **Written mostly in:** C

```
array([[0.8249888 , 0.24979155, 0.79719817, 0.0035682 , 0.00815663],
       [0.59428351, 0.5941183 , 0.63903329, 0.68025011, 0.69124751],
       [0.97160135, 0.75746315, 0.81417836, 0.2592979 , 0.51379441],
       [0.45281654, 0.49509363, 0.40326315, 0.11308091, 0.78175295],
       [0.61679325, 0.35484431, 0.34741814, 0.66497307, 0.06190098],
       [0.80918979, 0.19280099, 0.79689589, 0.36151733, 0.67275732],
       [0.22898206, 0.96506853, 0.9395566 , 0.83567071, 0.90740995],
       [0.5153297 , 0.73086199, 0.62606757, 0.20651277, 0.3589751 ],
       [0.83653033, 0.758272  , 0.35327387, 0.26504777, 0.37401509],
       [0.20079886, 0.25584164, 0.20877105, 0.37039681, 0.29956502],
       [0.37167533, 0.93510168, 0.77947436, 0.71571247, 0.40358436],
       [0.10809994, 0.54756469, 0.7417065 , 0.25672907, 0.0192784 ],
       [0.88182689, 0.66561254, 0.99875671, 0.42072552, 0.38811864],
       [0.12863145, 0.11459206, 0.20615946, 0.04105515, 0.03424916],
       [0.77313463, 0.59808671, 0.2568987 , 0.45116242, 0.75258347],
       [0.5283624 , 0.37202166, 0.83716868, 0.15319497, 0.13126105],
       [0.98516634, 0.28416098, 0.98175643, 0.98217529, 0.34898673],
       [0.15650063, 0.40911476, 0.63216031, 0.4845381 , 0.34404839],
       [0.25099388, 0.84347889, 0.95057475, 0.26486295, 0.22221181],
       [0.59412643, 0.58056145, 0.3299675 , 0.46726547, 0.35558934]])
```

20 x 5 Numpy array

Pandas

- **Type:** Data Science Matrices (numbers and words)
- **Main structure:** Dataframes - similar to SQL Tables
- **Use Cases:** Groupby/sorting, reading/writing files from csv, Excel, Google Drive, SQL, etc.; regression, plotting
- **Written mostly in:** Python and Cython, with critical functions in C

	Order ID	Order Date	Product ID	Destination Port	Unit quantity	Weight
0	1.447296e+09	2019-12-31	1700106	PORT09	808	14.30
1	1.447158e+09	2019-12-31	1700106	PORT09	3188	87.94
2	1.447139e+09	2019-12-31	1700106	PORT09	2331	61.20
3	1.447364e+09	2019-12-31	1700106	PORT09	847	16.16
4	1.447364e+09	2019-12-31	1700106	PORT09	2163	52.34

Pandas Dataframe

Introduction to Scikit-learn

Scikit-learn

- Machine learning library Built on top of NumPy library (and others)
- Data Preprocessing: Feature Extraction, Normalization
- Supervised Learning: Regression, Classification, Neural Networks
- Unsupervised Learning: Clustering, Dimensionality Reduction

Introduction to Scikit-learn

Scikit-learn

- Machine learning library Built on top of NumPy library (and others)
- Data Preprocessing: Feature Extraction, Normalization
- Supervised Learning: Regression, Classification, Neural Networks
- Unsupervised Learning: Clustering, Dimensionality Reduction

Typical Workflow

1) Read in Data

2) Transform Data

3) Apply Model

4) Analyze results

Introduction to Scikit-learn

Scikit-learn

- Machine learning library Built on top of NumPy library (and others)
- Data Preprocessing: Feature Extraction, Normalization
- Supervised Learning: Regression, Classification, Neural Networks
- Unsupervised Learning: Clustering, Dimensionality Reduction

Typical Workflow

1) Read in Data

2) Transform Data

3) Apply Model

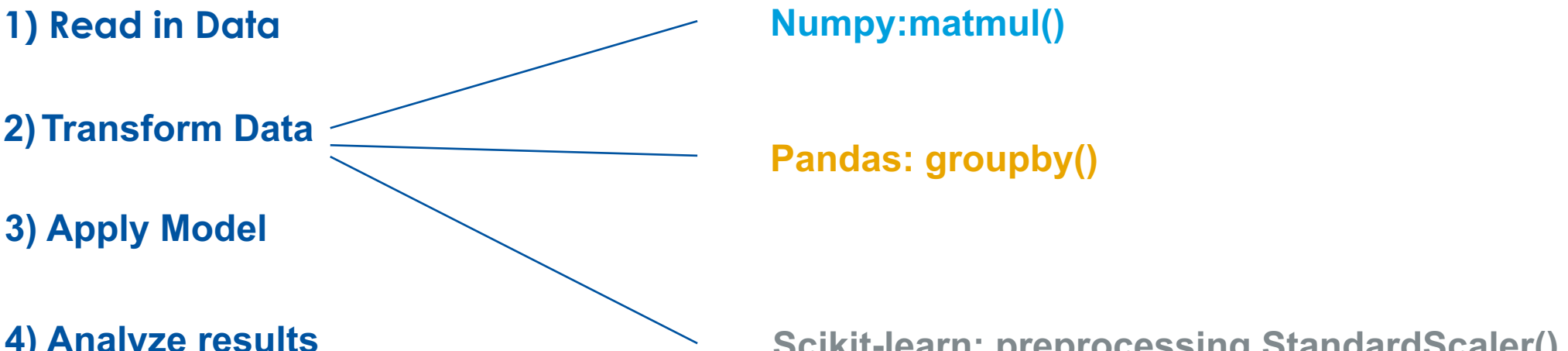
4) Analyze results

Introduction to Scikit-learn

Scikit-learn

- Machine learning library Built on top of NumPy library (and others)
- Data Preprocessing: Feature Extraction, Normalization
- Supervised Learning: Regression, Classification, Neural Networks
- Unsupervised Learning: Clustering, Dimensionality Reduction

Typical Workflow



Introduction to Scikit-learn

Scikit-learn

- Machine learning library Built on top of NumPy library (and others)
- Data Preprocessing: Feature Extraction, Normalization
- Supervised Learning: Regression, Classification, Neural Networks
- Unsupervised Learning: Clustering, Dimensionality Reduction

Typical Workflow

1) Read in Data

2) Transform Data

3) Apply Model ——————— **Scikit-learn: LinearRegression()**

4) Analyze results

Introduction to Scikit-learn

Scikit-learn

- Machine learning library Built on top of NumPy library (and others)
- Data Preprocessing: Feature Extraction, Normalization
- Supervised Learning: Regression, Classification, Neural Networks
- Unsupervised Learning: Clustering, Dimensionality Reduction

Typical Workflow