DATA SCIENCE DEVELOPMENT TOOLS

PART 1: PYTHON, NUMPY

TOOLS FOR DATA MINING

- Start with Pickaxe and Shovel
- Move to jack hammer and other automated tools
- Python, SQL, R most common languages
- NumPy, Pandas, Matplotlib, Seaborn, Scikit-Learn
- Tableau, PowerBI
- AWS SageMaker, Dataiku, DataRobot, DataBricks

COMMON DATA SCIENCE PACKAGES FOR PYTHON

CS 180

Import name	Common alias	Description
`numpy`	`np`	NumPy includes functions and classes that aid in numerical computation. `NumPy` is used in many other data science packages.
`pandas`	`pd`	pandas provides methods and classes for tabular and time-series data.
`sklearn`	`sk`	scikit-learn provides implementations of many machine learning algorithms with a uniform syntax for preprocessing data, specifying models, fitting models with cross-validation, and assessing models.
`matplotlib.pyplot`	`plt`	matplotlib allows the creation of data visualizations in Python. The functions mostly expect `NumPy` arrays.
`seaborn`	`sns`	seaborn also allows the creation of data visualizations but works better with pandas DataFrame.
`scipy.stats`	`sp.stats`	SciPy provides algorithms and functions for computing problems that arise in science, engineering and statistics. `scipy.stats` provides the functions for statistics.
`statsmodels`	`sm`	statsmodels adds functionality to Python to estimate many different kinds of statistical models, make inferences from those models, and explore data.

BYU

NUMPY

- **Spelled: NumPy, Pronounced “num-pie”**
- **What is NumPy?** NumPy (Numerical Python) is the fundamental package for scientific computing with Python. It provides a high-performance multidimensional array object and tools for working with these arrays.
- **Why NumPy?**
  - **Speed:** NumPy arrays are more efficient and faster than Python lists for numerical operations, as they are implemented in C.
  - **Functionality:** It provides a rich set of functions for linear algebra, Fourier transforms, and random number generation.
  - **Foundation:** Many other data science libraries like Pandas, SciPy, and Scikit-learn are built on top of NumPy.

BYU

NUMPY ARRAY FUNCTIONS

Function	Parameters	Description

`array()`	`object`
	`dtype=None`
	`ndim=0`
		Returns an array constructed from **`object`**. **`object`** must be a scalar or an ordered container, such as tuple or list. The array element type is inferred from **`object`** unless a **`dtype`** is specified. **`ndim`** is the minimum number of array dimensions.

`delete()`	`arr`
	`obj`
	`axis=None`
		Deletes a slice of input array **`arr`**. **`axis`** is the axis along which to remove a slice. **`obj`** is the index of the slice along the axis.

`full()`	`shape`
	`fill_value`
	`dtype=None`
		Returns an array filled with **`fill_value`**. The **`shape`** tuple specifies array shape. **`dtype`** specifies the array type. If **`dtype=None`**, the type is inferred from **`fill_value`**.

`insert()`	`arr`
	`obj`
	`values`
	`axis=None`
		Inserts array **`values`** to input array **`arr`**. **`axis`** is the axis along which to insert. **`obj`** is the index before which **`values`** is inserted.

`zeros()`	`shape` `dtype=float`	Returns an array filled with zeros. The **`shape`** tuple specifies array shape. **`dtype`** specifies the array type.
`ones()`	`shape` `dtype=None`	Returns an array filled with ones. The **`shape`** tuple specifies array shape. **`dtype`** specifies the array type. If **`dtype=None`**, the type is float64.
`sort()`	`a` `axis=-1`	Sorts array **`a`** along **`axis`**. The default **`axis=-1`** sorts along the last axis in **`a`**. **`axis=None`** flattens **`a`** before sorting.

BYU