

# Data Science Development Tools

Part 1: Python, NumPy

# Announcements

- Updated Discord Link
- Update TA Contact Info
- Updated TA Hours
- Chap 2 Reading - Part 1 - Due Jan 17

Class #	Week #	Month	Date	Topic	Reading	Labs
1	1	Jan	8	Welcome, Introduction, Course Objectives, DS Lifecycle	Chapter 1 Intro DS	Lab 1: Colab Set Up, GitHub
2	2	Jan	13	Python setup, Google colab, Github		
3	2	Jan	15	NumPy, Vectorization	Chapter 2.1-2.5 Python	Lab 2: Vectorization
4	3	Jan	20	Pandas, Matplotlib, Seaborn		
5	3	Jan	22	Data Cleaning and Preparation	Chapter 2.6-2.8 Python	Lab 3: NumPy, Pandas
6	4	Jan	27	Data Acquisition, ETL, Populations, Sampling	Chapter 3 Data Prep	
7	4	Jan	29	Descriptive Statistics	Chapter 4 Prob & Stat	Lab 4: Data Preparation
8	5	Feb	3	Exploratory Data Analysis (EDA)		
9	5	Feb	5	Principles of Data Visualization	Chapter 5 EDA	
10	6	Feb	10	Data management - databases, SQL queries		
11	6	Feb	12	More SQL Features, Joins	Chapter 7 DB and SQL	
12	7	Feb	17	MONDAY SCHEDULE, NO CLASS		
13	7	Feb	19	SQLite		
14	8	Feb	24	MIDTERM REVIEW		
15	8	Feb	26	MIDTERM		
16	9	Mar	3	Overview of ML		
17	9	Mar	5	Unsupervised Learning- Kmeans	Chapter 8 Unsupervised Learning	
18	10	Mar	10	Unsupervised Learning- Hierarchical, DBSCAN		
19	10	Mar	12	Supervised Learning: Part 1	Chapter 9 Supervised Learn	Lab 8: Cluster Analysis
20	11	Mar	17	Supervised Learning: Part 2		
21	11	Mar	19	Evaluation of models, comparing performance	Chapter 10 Decision Trees	Lab 9: ML Classification/Regression
22	12	Mar	24	Feature Importance with RF and Logistic Regression		
23	12	Mar	26	ANN, Multi-Layer Perceptron, Backpropagation	Chapter 12 Eval	New Lab?
24	13	Mar	31	Deep Learning		
25	13	Apr	2	GenAI - Introduction	Chapter 13 ANN	Lab 10: MLP and Backpropagation



You are here

# Tools for Data Mining

- Start with Pickaxe and Shovel
- Move to jack hammer and other automated tools
- **Languages** - Python, SQL, R
- **Editors** — Jupyter, Colab, VSCode, PyChar etc.
- **Libraries** - NumPy, Pandas, Matplotlib, Seaborn, Scikit-Learn
- **Business Intelligence Platforms** - Tableau, PowerBI
- AWS QuickSight/SageMaker, Dataiku, DataRobot, DataBricks



# Common Data Science Packages for Python

CS 180

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

# 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.
- Originally Developed by **Travis Oliphant** - Assistant Professor 2001-2007

# NumPy Array Functions

Function	Parameters	Description
<code>array()</code>	<code>object</code> <code>dtype=None</code> <code>ndim=0</code>	Returns an array constructed from <code>object</code> . <code>object</code> must be a scalar or an ordered container, such as tuple or list. The array element type is inferred from <code>object</code> unless a <code>dtype</code> is specified. <code>ndim</code> is the minimum number of array dimensions.
<code>delete()</code>	<code>arr</code> <code>obj</code> <code>axis=None</code>	Deletes a slice of input array <code>arr</code> . <code>axis</code> is the axis along which to remove a slice. <code>obj</code> is the index of the slice along the axis.
<code>full()</code>	<code>shape</code> <code>fill_value</code> <code>dtype=None</code>	Returns an array filled with <code>fill_value</code> . The <code>shape</code> tuple specifies array shape. <code>dtype</code> specifies the array type. If <code>dtype=None</code> , the type is inferred from <code>fill_value</code> .
<code>insert()</code>	<code>arr</code> <code>obj</code> <code>values</code> <code>axis=None</code>	Inserts array <code>values</code> to input array <code>arr</code> . <code>axis</code> is the axis along which to insert. <code>obj</code> is the index before which <code>values</code> is inserted.
<code>zeros()</code>	<code>shape</code> <code>dtype=float</code>	Returns an array filled with zeros. The <code>shape</code> tuple specifies array shape. <code>dtype</code> specifies the array type.
<code>ones()</code>	<code>shape</code> <code>dtype=None</code>	Returns an array filled with ones. The <code>shape</code> tuple specifies array shape. <code>dtype</code> specifies the array type. If <code>dtype=None</code> , the type is float64.
<code>sort()</code>	<code>a</code> <code>axis=-1</code>	Sorts array <code>a</code> along <code>axis</code> . The default <code>axis=-1</code> sorts along the last axis in <code>a</code> . <code>axis=None</code> flattens <code>a</code> before sorting.

# Upcoming Assignments

- Reading: 2.1 – 2.5
- DS Lab 2: Vectorization due Jan 17<sup>th</sup> 11:59 pm