Last Updated: 9/11/2022

Data 601: Introduction to Data Science Reference Guide

Course Main Tools and Important References

- Jupyter Notebook (https://jupyter.org/) is a web application for creating and sharing computational documents. It offers a simple, streamlined, document-centric experience. Various ways to run the Jupyter Notebook environment are shown in
- Python Documentation (https://docs.python.org/3/)

Once Jupyter is installed it can be run in multiple ways as described in

Table 1: Ways to Run Jupyter Notebooks

Method to run Jupyter Notebooks	Comments
Windows Jupyter Notebook Launcher	Launch the Windows Desktop application directly. Instructor's preferred method.
Anaconda Navigator Jupyter Notebook	Same as above but adds an extra step. Launch from the Anaconda Navigator application.
Integrated Development Environment (IDE)	IDE's such as Visual Studio Code allow to use the Jupyter Notebook outside of the Web-browser environment. It also allows seamless transition when working on different codes in the IDE as well as managing different files.
Google Colab Research (https://colab.research.google.com/)	May not be an option when using internal, non-public data. Check with your organization.

The recommended approach is to use the tools that your work team is using. Integrating into your team will be critical and will support better communication, collaboration, debugging and exchange of information.

Table 2: Important Python Data Science References

Title	Description	Link
Python Documentation	Python Documentation	https://www.python.org/doc/ and https://docs.python.org/3/
Anaconda Included Packages List	Anaconda Distribution List of Packages	https://docs.anaconda.com/anaconda/packages/pkg-docs/

This table has a list of common and important Python data science packages and libraries. It also includes the link to the libraries Python Package Index (https://pypi.org/) as well as a short description. The table includes those packages and libraries included within Anaconda distribution. Alternatively, Google Collaboratory can call each library. Many of these will be discussed in the Data601 class while others are for reference for future classes or work projects.

Note that libraries could have overlapping scope, capabilities, and features. For example, a linear regression model can be created with Numpy, scikit-learn, or sciPy. Deciding which library to use will depend on user familiarity with the module, speed, limitation and other considerations.

Table 3: Common and Important Python Data Science Packages and Libraries

Library/ Package Title	Included in Anaconda	Description and main use
PIP	Yes	Pip is the package installer for Python.
Pandas	Yes	Pandas is a Python package that provides fast, flexible, and expressive data structures designed to make working with "relational" or "labeled" data both easy and intuitive.
Numpy	Yes	Numpy is a powerful N-dimensional array object, sophisticated (broadcasting) functions, tools for integrating C/C++ and Fortran code, useful linear algebra, Fourier transform, random number capabilities, and much more.
Matplotlib	Yes	Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python.
Plotly	Yes	Plotly.py is an interactive, open-source, and browser-based graphing library for Python.
Seaborn	Yes	Seaborn is a Python visualization library based on matplotlib. It provides a high-level interface for drawing attractive statistical graphics.
Re (Regex or Regular Expressions)	Yes	This package pertains specifically to regular expressions embedded inside Python and compiled with Python's [`re`] (https://docs.python.org/3/library/re.html) module.
OS	Yes	This module provides a portable way of using operating system dependent functionality.
Requests	Yes	Requests is a simple, yet elegant, HTTP library.
BeautifulSoup4	Yes	Beautiful Soup is a library that makes it easy to scrape information from web pages. It sits atop an HTML or XML parser, providing Pythonic idioms for iterating, searching, and modifying the parse tree.
Scrapy	Yes	Scrapy is a fast high-level web crawling and web scraping framework, used to crawl websites and extract structured data from their pages.
Selenium	No	Python language bindings for Selenium WebDriver. The selenium package is used to automate web browser interaction from Python. Can be used for dynamic JavaScript based websites.
Scikit-learn	Yes	Scikit-learn is a Python module for machine learning built on top of SciPy and is distributed under the 3-Clause BSD license.
Scipy	Yes	SciPy is open-source software for mathematics, science, and engineering. The SciPy library depends on NumPy, which provides convenient and fast N-dimensional array manipulation.

Last Updated: 9/11/2022

Library/ Package Title	Included in Anaconda	Description and main use
Statsmodels	Yes	Statsmodels is a Python package that provides a complement to scipy for statistical computations including descriptive statistics and estimation and inference for statistical models.
spaCy	No	SpaCy is a library for advanced Natural Language Processing in Python and Cython.
NLTK	Yes	The Natural Language Toolkit (NLTK) is a Python package for natural language processing.
Gensim	Yes	Gensim is a Python library for topic modelling, document indexing and similarity retrieval with large corpora.
SQLite	Yes	SQLite is a C library that provides a lightweight disk-based database that doesn't require a separate server process and allows accessing the database using a nonstandard variant of the SQL query language.
Openpyxl	Yes	Openpyxl is a Python library to read/write Excel 2010 xlsx/xlsm/xltx/xltm files.
Networkx	Yes	NetworkX is a Python package for the creation, manipulation, and study of the structure, dynamics, and functions of complex networks.
Ipywidgets	Yes	Ipywidgets, also known as jupyter-widgets or simply widgets, are interactive HTML widgets for Jupyter notebooks and the IPython kernel.
Panel	Yes	Panel provides tools for easily composing widgets, plots, tables, and other viewable objects and controls into custom analysis tools, apps, and dashboards
Flask	Yes	Flask is a lightweight WSGI web application framework. It is designed to make getting started quick and easy, with the ability to scale up to complex applications.
Django	No	Django is a high-level Python web framework that encourages rapid development and clean, pragmatic design.
TensorFlow	No	TensorFlow is an open-source software library for high performance numerical computation (https://www.tensorflow.org/)
TensorFlow Keras	No	Keras is TensorFlow's module for deep learning.
Pytorch	No	PyTorch is a Python package that provides two high-level features: (1) Tensor computation (like NumPy) with strong GPU acceleration and (2) Deep neural networks built on a tape-based autograd system.

The table below has websites that have various sample datasets.

Table 4: Sites with Sample Datasets

Title	Description	Link	Comment
Github	Used by many developers to share code and collaborate.	https://github.com/	Class code is hosted in this website.
U.S. Government's Open Data	Data, tools, and resources to conduct research, develop web and mobile applications, design data visualizations, and more.	https://data.gov/	Real world datasets that in some cases may be extremely large.
Kaggle	Host data science competitions, datasets, Jupyter notebooks, etc.	https://www.kaggle.com/	Good source for datasets and Jupyter Notebooks.
Google Research Data	Google Dataset Search Engine	https://datasetsearch.research.google.com/	Search engine for datasets. May send you to data in other websites in this table.
World Bank Data	Data collections by the World Bank	https://datacatalog.worldbank.org/	Also has collections from various sources.
Maryland Government Data	MD Open Data Website.	https://opendata.maryland.gov/	Real world datasets from Maryland's State Government

The table contains reference websites that include sample galleries, how to use various plotting libraries with examples and articles on visualizations.

Table 5: Sites with Visualization Examples

Title	Description	Link	Comments
MatPlotLib Example Gallery	Python Package	https://matplotlib.org/stable/gallery/index.html	Included in Anaconda.
Seaborn Example Gallery	Python Package	https://seaborn.pydata.org/examples/index.html	Included in Anaconda.
Plotly Example Gallery	Python Package	https://plotly.com/python/	Included in Anaconda.
D3.JS	Website Plotting Library	https://d3-graph-gallery.com/	Used to deploy dynamic website plots.
Five Thirty Eight	Website	https://fivethirtyeight.com/	Excellent Website that provides great examples on the capabilities of dashboards and data analytic visualizations.

The table below contains various websites with news, blogs and articles related to data analytics, data science, artificial intelligence, machine learning, natural language processing among other related topics.

Table 6: Data Science News, Blogs and Articles

Title	Description	Link
50 years of data science	Paper	https://courses.csail.mit.edu/18.337/2015/docs/50YearsDataScience.pdf
A Very Short History Of Data Science	Article	https://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/
Medium	Subscription based website with limited amount of free articles per month.	https://medium.com/
KD Nuggets Website	Website	https://www.kdnuggets.com/
Y Combinator News	Website	https://news.ycombinator.com/
Hacker Noon Website	Website	https://hackernoon.com/
Reddit Data Science Blogs	Reddit Website Blogs	https://www.reddit.com/r/datascience/
Data Elixir Newsletter	Website	https://dataelixir.com/newsletters/
Inside Big Data Website	Website	https://insidebigdata.com/
AI Google Blog	Blog	https://ai.googleblog.com/
Machine Learning Mastery Website	Website	https://machinelearningmastery.com/

Table 7: Meetups

Title	Link
Data Science Meetup	https://www.meetup.com/topics/data-science/
Data Works Meetup	https://www.meetup.com/DataWorks/
Statistical Seminars Meetup	https://www.meetup.com/Statistical-Seminars-DC/

The following table contains various references, standards and frameworks related to ethics considerations.

Table 8: AI Standards, Frameworks and Ethics Considerations

Organization	**Description**	**Link**
National Institute of Standards and Technology (NIST)	NIST artificial intelligence references.	https://www.nist.gov/artificial-intelligence
NIST	NIST artificial intelligence (AI) risk management framework to better manage risks to individuals, organizations, and society associated with AI.	https://www.nist.gov/itl/ai-risk-management-framework
U.S. Government Accountability Office (GAO)	GAO AI framework report identifies key accountability practices centered around the principles of governance, data, performance, and monitoring to help federal agencies and others use AI responsibly.	https://www.gao.gov/products/gao-21-519sp
Institute of Electrical and Electromechanical Engineers (IEEE)	IEEE AI website	https://standards.ieee.org/initiatives/artificial-intelligence-systems/
IEEE 7000™-2021	IEEE 7000™-2021 integrates ethical and functional requirements to mitigate risk and increase innovation in systems engineering product design.	https://standards.ieee.org/news/2021/ieee-7000/