# CSI500: Data Science Tools – Introduction to Jupyter Notebooks

|  |  |  |  |
| --- | --- | --- | --- |
| **Learning Objectives** | **Discussion(s)** | **Assignment(s)** | **Learning Resources + Media** |
| After this lesson, you will be able to …   * Import a classic machine learning data science data set into a notebook * Use Python data science tools to perform descriptive statistics on the data set * Use Python tools to create scientific plots and graphs from the data in the notebook * Answer botanical questions by using descriptive statistical results and scientific graphs rendered in the notebook | Previously, we have used Python in the Spyder IDE. For this lesson, we will explore using Python with Jupyter Notebooks.  We want to answer these questions:  • What are the trade-offs between Python in Spyder versus Python in Jupyter Notebooks?  • When is one tool better than the other?  • How can Jupyter Notebooks help me explore data sets?  • How can I make my visualizations more convincing for the reader? | Through the Iris dataset, you will examine biological features (specifically sepal and petal lengths) among three different species of common Iris flowers: Iris *setosa,* Iris *versicolour, and* Iris *virginica.* This data set was originally developed by the prominent British statistician R.A. Fisher, and is considered a classic in the machine learning literature.  **To prepare for this assignment**:   * Download and install the Anaconda Python distribution (which should include Jupyter notebooks by default) * Download the Iris flower dataset   **Assignment**  **Part 1 – Data Import**  Import the Jupyter notebook provided by the instructor.  Execute the code block to import the Iris data set.  Answer the questions about the basic data set “shape” (number of rows, number of columns, column names, Python data type(s) used)  **Part 2 – Descriptive Statistics?**  Run the code blocks for descriptive statistics.  Identify basic statistical measures for the various observations  Mean, standard deviation  **Part 3 – Scientific Visualization**  Run the code blocks to generate plots using ggplot()  Develop scatter plots of the data observations  Use color coding and shading to make the data easier to view  **Part 4 – Analysis**  Based on the results of Part 2 and Part 3, answer questions about the various species of Iris.  In particular, determine if it is possible to automatically classify an observation based on one or more features from the data set?  Explain why or why not. | • Readings: Websites, lecture notes  • Video Lectures  • Worksheet  Dataset: <https://en.wikipedia.org/wiki/Iris_flower_data_set>  Built-into the Python scikit-learn package. Instantiated using the “load\_iris()” command from within the Notebook. |