Setting up Python for Intro to Data Science

After going through this document you will have set up a working Python environment to use for your own data science projects. Make sure to read the document in its entirety before taking any steps and make sure you understand each step clearly. If you get stuck on any step our coaches will be happy to assist you!

Course Files:

Download the course files and save them to any directory you wish!

In the course files you should have folders for the programming assignments in lesson 2 through 5 and the folders for each of the 5 course projects.

Install Anaconda:

If you are an absolute beginner I strongly suggest you install the Anaconda development environment. This will install all of the necessary packages (with the exception of pandasql and ggplot, which become relatively easy to install afterwards). Make sure to go through the Anaconda install guide.

If you plan to use a python IDE you should check out this guide for setting up Anaconda on your IDE.

After you have Anaconda set up you will need to install pandasql and ggplot. Current distributions of Anaconda include the package manager pip as part of the installation; you can also follow this link to get instructions to set up pip if necessary. Once you have pip

installed, use `pip install pandasql` and `pip install ggplot`. Alternatively, you can download pandasql and ggplot and run the setup.py file that is included with them in Anaconda.

Install Python:

If you've followed the previous section for installing Anaconda, you shouldn't need to perform the steps in this and the following sections. If you'd like to install Python on its own, from the Python website download and install the latest version of Python 2.7.x. We strongly suggest downloading Python 2.7.x and avoid using Python 3.x.x at this time. While Python 3 is the future of Python many of the libraries used in this course will not work with Python 3.

Once you've successfully installed Python you should also have IDLE which is the primary Python IDE you can use. (Note: Feel free to use Vim, Sublime Text, GEdit, Notepad, etc. if you are comfortable developing in these environments.)

You can learn more about using IDLE here. If you'd prefer to use your OS's console or terminal to compile and run python code look no further than Python's own online documentation.

Installation of Required Packages:

Install packages in the following order Numpy, Scipy, Matplotlib, Pandas, Statsmodels, ggplot, pandasql.

Windows
(Note: I **highly** recommend Anaconda for Windows users):
Pandas, Numpy, Scipy, Matplotlib, and Statsmodels can all be found at Christopher Gohlke's Unofficial Windows Binaries for Python Extenstion Packages. Just download the executable and run the file on your PC and these packages will become available in your python distribution. (Note: Make sure you download the 2.7.x versions matching your version of Python.)

Unfortunately getting ggplot and pandasql isn't as easy for Windows. You will have to set up pip. (Note you will have to set up the python interpreter for console use). Once pip is installed it is easy enough to run the commands: `pip install pandasql` and `pip install ggplot`

Mac OSX:

If you are familiar with pip, you can follow the steps outlined in the Windows section to install the necessary packages. Alternatively, you can visit the Python Package Index to search for each of the necessary packages. Download the .tar.gz file and extract each package. Once extracted there should be a setup.py file. Run this file and the package will install to your Python distribution.