

Data Science Essentials

Lab Setup Guide

Overview

This guide takes you through the steps to create an environment for performing the data science experiments described in this repository.

To prepare the lab environment, you must perform the following tasks:

1. Create an Azure ML account
2. Download and extract the lab files
3. Install Microsoft R Open and RStudio
or
4. Install Python Anaconda

What You'll Need

To perform the setup tasks, you will need the following:

- A Windows, Linux, or Mac OSX computer.
- A web browser and Internet connection.

Create an Azure ML Account

Azure ML offers a free-tier account, which you can use to complete the labs in this repository.

Note: A free Azure ML workspace is not the same as a Microsoft Azure trial subscription!

Sign Up for a Microsoft Account and a Free Azure ML Workspace

1. If you do not already have a Microsoft account, sign up for one at <https://signup.live.com/>.
2. Browse to <http://aka.ms/edx-dat203.1x-aml> and click **Get Started Now**. Then follow the instructions to sign up for a free Azure ML workspace. If prompted, sign in with your Microsoft account credentials.

Note: Your free-tier Azure ML workspace allows you unlimited access, with some reduced capabilities compared to a full Microsoft Azure subscription. Your experiments will only run at low priority on a single processor core. As a result, you will experience some longer wait times. However, you have full access to all features of Azure ML.

Download the Lab Files

All code and data files you will need are contained in a zip file. Follow these steps to download and install the lab files.

1. Download the lab files from <http://aka.ms/edx-dat203.1x-labfiles>.

2. Extract the downloaded zip file to a convenient folder on your local computer.

Install Microsoft R Open and RStudio

R is a programming language for conducting statistical analysis processes and visualizing data. You will create R code for some of the labs in this course.

Note: For this course, you can choose to complete programming exercises in Python or R (or both if you are truly ambitious!). If you plan to use R, complete this procedure to install the R runtime and development tools. If you do not plan to use R, you can skip this section and go to the procedure for installing Anaconda Python.

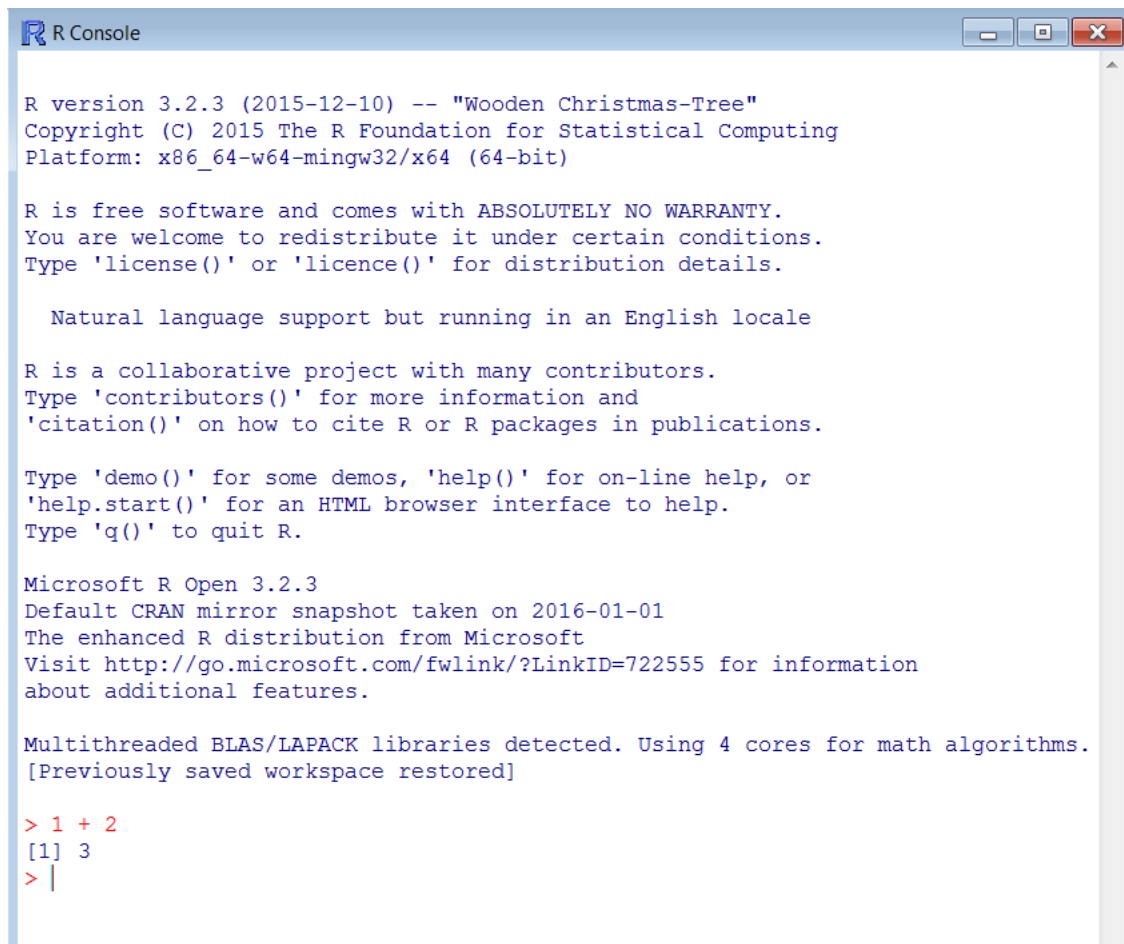
Install Microsoft R Open

Note: Microsoft R Open is an enhanced version of R. Microsoft R Open uses the multi-core MKL math libraries to enhance performance. However, if are using GNU R downloaded from CRAN, you can continue to do so for this course.

1. In a web browser, navigate to <http://aka.ms/edx-dat203.1-R>.
2. Select the **Microsoft R Open** download for your operating system and following the installation directions.

Note: If you are using Mac OS X, you can skip the following step as the MKL math libraries are already included.

3. When the Microsoft R Open installation has finished, select the **MKL** math library download for your operating system and follow the installation directions.
4. Verify your installation by starting Microsoft R Open, from the desktop icon, and entering a simple R expression such as $1 + 2$ (which should produce the result 3) in the console as shown in the following image.



```
R Console

R version 3.2.3 (2015-12-10) -- "Wooden Christmas-Tree"
Copyright (C) 2015 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

Microsoft R Open 3.2.3
Default CRAN mirror snapshot taken on 2016-01-01
The enhanced R distribution from Microsoft
Visit http://go.microsoft.com/fwlink/?LinkID=722555 for information
about additional features.

Multithreaded BLAS/LAPACK libraries detected. Using 4 cores for math algorithms.
[Previously saved workspace restored]

> 1 + 2
[1] 3
> |
```

5. Close R.

Install RStudio

1. In a web browser, navigate to <https://www.rstudio.com/products/rstudio/download/>.
2. Run the installer for your operating system (Windows, MacOSX, Ubuntu, or Fedora) to install RStudio.
3. To verify installation, start RStudio, and look at the console which should resemble the following image:

```
Console C:/Users/Steve/Dropbox/Azure ML/Data Science Essentials/Mod 4/

R version 3.2.3 (2015-12-10) -- "Wooden Christmas-Tree"
Copyright (C) 2015 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

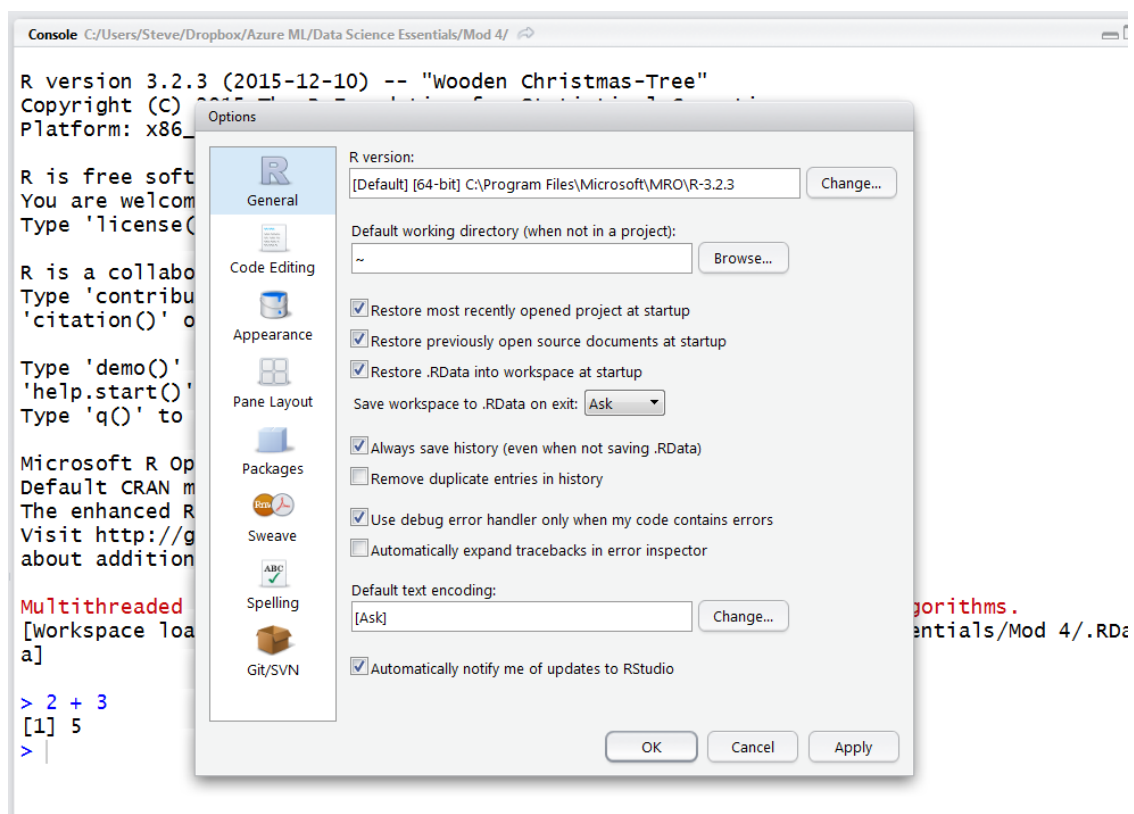
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

Microsoft R Open 3.2.3
Default CRAN mirror snapshot taken on 2016-01-01
The enhanced R distribution from Microsoft
Visit http://go.microsoft.com/fwlink/?LinkID=722555 for information
about additional features.

Multithreaded BLAS/LAPACK libraries detected. Using 4 cores for math algorithms.
[Workspace loaded from C:/Users/Steve/Dropbox/Azure ML/Data Science Essentials/Mod 4/.RDa
a]

> 2 + 3
[1] 5
> |
```

4. Verify that RStudio is configured to use the current version Microsoft R Open, by noting the version of Microsoft R Open displayed on the console. (The current version is 3.2.3).
5. If your configuration is not correct select **Global Options** on the **Tools** menu and set the path to the directory where you installed Microsoft R Open, as shown in the following image.



- When you have verified installation and configuration, close RStudio.

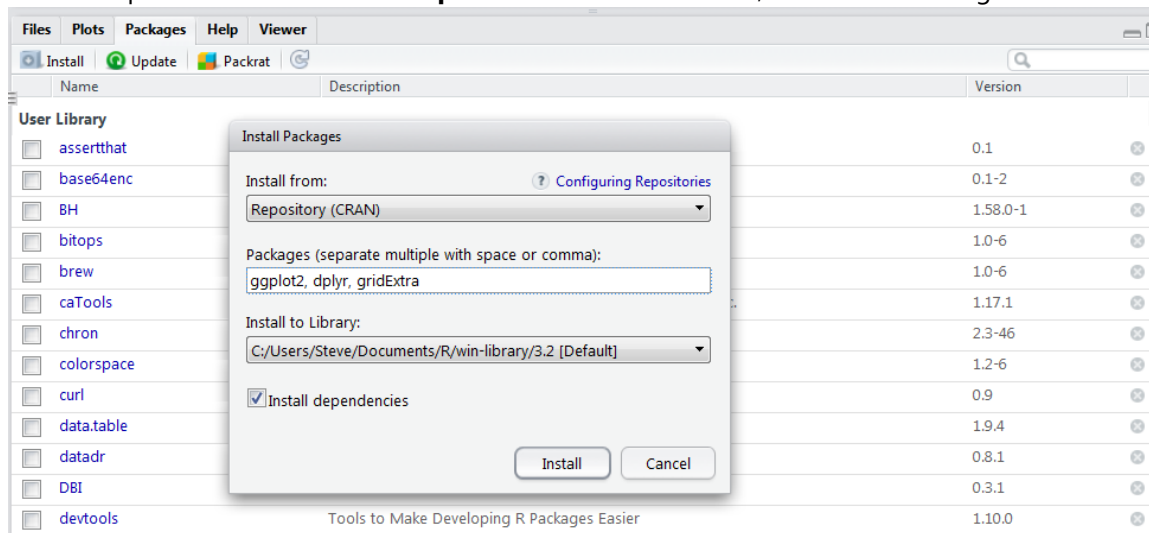
Install R packages

Several of the labs for this course require R packages that are not installed by default. These packages include:

- ggplot2
- gridExtra
- dplyr

Follow these instructions to install these packages:

- In RStudio, locate the pane with **Packages** tab and click on it.
- Click the **Install** icon.
- In the Packages text box of the dialog, type '**ggplot2, dplyr, gridExtra**', ensuring the names are comma separated and the **Install dependencies** box is checked, as shown in the figure.



- Click **Install**. Expect a great deal of text to appear on the console. Watch for error messages.
- At the console prompt in RStudio, type the following commands to test loading the packages:

```
library(ggplot2)
library(dplyr)
library(gridExtra)
```

Install Python Anaconda

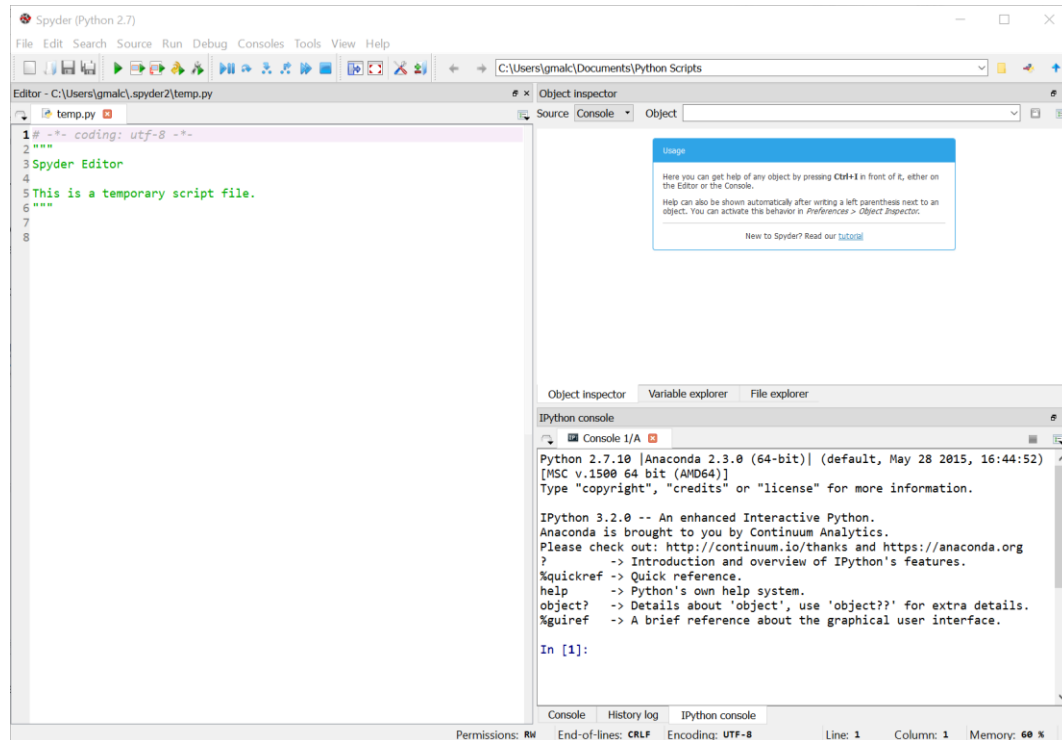
Python Anaconda is a distribution of Python that includes the Spyder Integrated Development Environment (IDE), which you will use to create Python code in the labs for this repository.

Note: In this repository, you can choose to complete programming exercises in Python or R (or both). If you plan to use Python, complete this procedure to install the Python runtime and development tools. If you do not plan to use Python, you can skip this procedure.

Install the Python Anaconda Distribution

Note: Microsoft Azure Machine Learning uses a version of the Anaconda Python distribution. However, if you are used to using another scientific Python distribution, such as Canopy, it is likely that you will be able to perform all of the exercises for this course using that distribution.

1. In a web browser, navigate to <http://continuum.io/downloads>.
2. Choose the installer for your operating system (Windows, Apple Macintosh, or Linux).
3. Complete the installation process for **Python 2.7** (Not Python 3.X).
4. After installation is complete, verify the installation by starting Spyder, which should look similar to the following image:



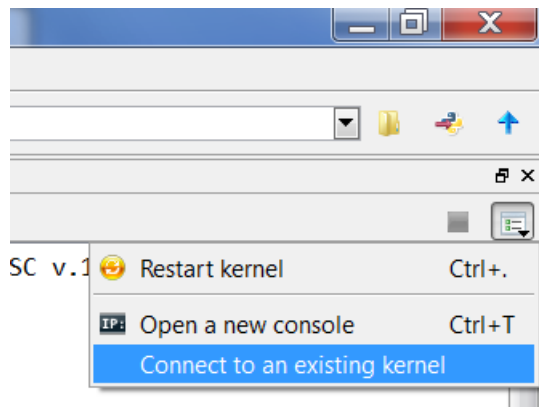
5. Close Spyder.

Verify Connection to the IPython kernel

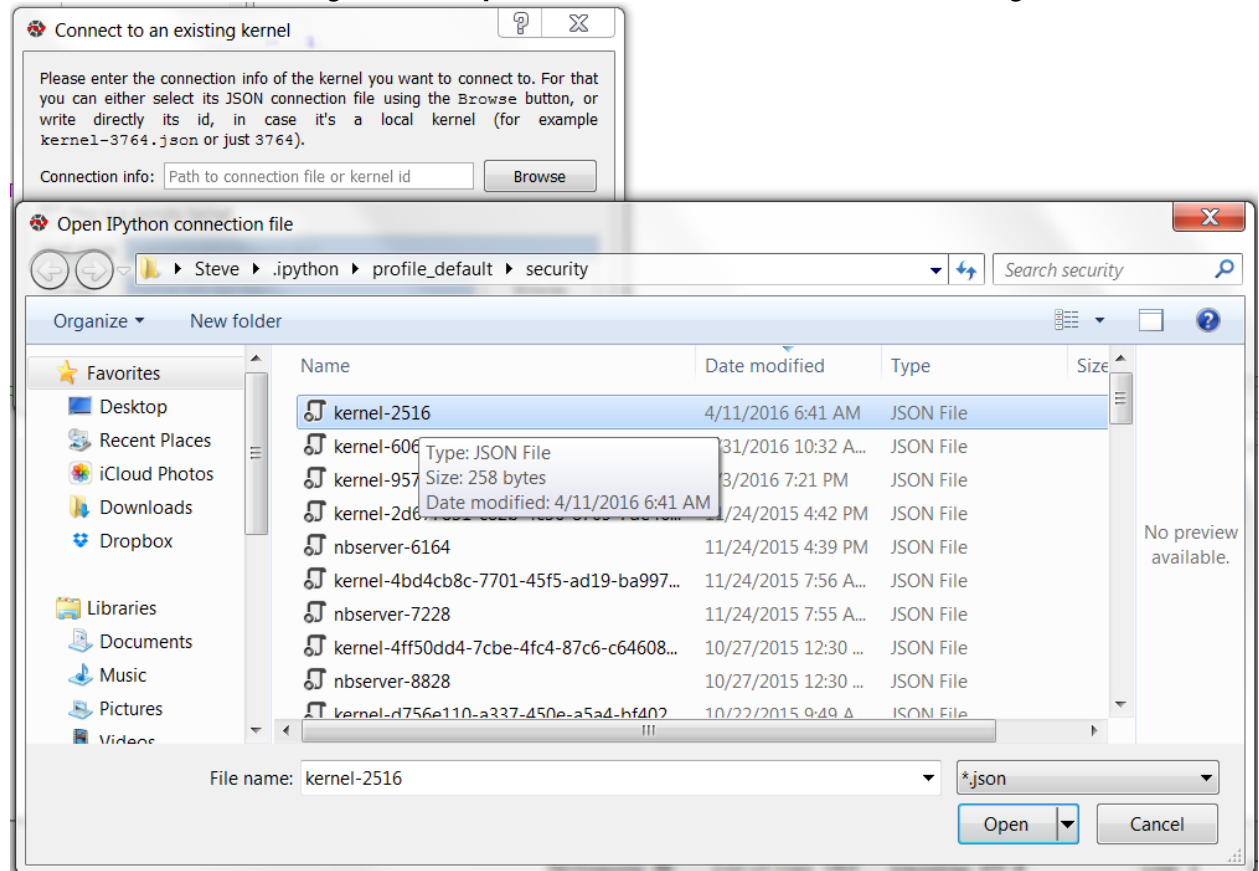
In some cases, when you first start Spyder, you may find that your IPython console does not respond to the commands you type at the prompt. If this occurs, wait a while to allow IPython kernel to start.

If the IPython console continues to be unresponsive, you may need to manually connect to an IPython kernel manually. Follow the steps below to connect a console to an IPython kernel.

1. From the kernel icon in the upper right above the IPython console window, select **Connect to an existing kernel** as shown in the figure below:



2. Select **Browse** on the dialog and then **Open** the first kernel on the list shown in the figure below:



3. Click OK, to complete selection of the Python kernel for your IPython session.

Summary

By completing the tasks in this setup guide, you have prepared your environment for the labs in this course. Now you're ready to start learning how to build data science and machine learning solutions.