Chapter 2: Building Your Analytics Toolbox: A Primer on Using R & Python For Security Analysis

If you add a little to a little and do this often, soon the little will become great.

Hesiod

Before we jump right into the various use cases in the book it’s important to ensure you at least have a basic familiarity with the two most prominent languages featured in nearly all of the scenarios: Python (http://www.python.org/) and R (http://www.r-project.org/). While there are an abundance of tools available for data analysis, we feel these two provide virtually all the features necessary to help you go from data to discovery with the least amount impedance.

A sub-theme throughout the book, and the distilled process at the heart of security data science, is: **idea**, **exploration**, **trial** (and *error*) and **iteration**. It is ineffective at best to attempt to shoehorn this process into a traditional *edit*/*compile*/*run* workflow found in most traditional languages and development environments. The acts of performing data analysis tasks and creating informative visualizations are highly interactive and iterative endeavors. Despite all of their positive features, even standalone Python and R do not truly enable rich, dynamic interaction with code and data; but, when coupled with IPython (http://ipython.org/) and RStudio (http://www.rstudio.com/), respectively, they are both transformed into powerful exploration tools, enabling rapid development and testing of everything from gnarly data munging to generating sophisticated visualizations.

This chapter will: provide pointers to installation resources for each tool, introduce core features of each language and development environment and explain the structure of the examples you will find in the remaining chapters of the book.

Why Python; Why R; *And*, Why Both?

A discussion of which programming language is better than another for a certain set of tasks often turns (quickly) into a religious war of words that rarely wins converts and never becomes fully resolved. As a security data scientist, you will find that you do not have the luxury of language bias since there will be times that one language shines in one area while a different one shines in another, and you will need the skills of a diplomat to bring them both together to solve real problems.

We’ve honed in on both R/RStudio and Python/IPython/pandas in this book, as they are the two leading data analysis languages/environments with broad similarities but also with unique elements that make them work well for some tasks and not others. As we go into the rationale behind each choice and as you become proficient in one or both environments, do not lull yourself into a sense of complacency. A hallmark of a good data scientist is adaptability and you should be continually scouring the digital landscape for emerging tools that will help you solve problems. We’ll introduce you to some of these upstarts in Chapter 13.

Why Python?

Guido van Rossum created the Python programming language in December of 1989 to solve a problem. He and his colleagues needed a common way to orchestrate system administration tasks that could take advantage of specific features in the operating systems they were using at that time. While there were existing interpreted, administrator-friendly tools and languages availabl, none were designed (from Guido van Rossum’s point of view) with either the flexibility or extensibility features baked into the design principles of Python.

Pyhon’s flexibility and extensibility (and the fact that it was free as in both “speech” and “beer”) were especially appealing to the scientific, academic and industrial communities starting in the early 2000s, and innovators in these fields quickly adapted this general purpose programming language to their own disciplines to solve problems easier than—ostensibly—the domain-specific languages available at that time.

You would have to search long and hard to find a file-type Python cannot read, a database Python cannot access and an algorithm Python cannot execute. As you familiarize yourself with the language, Python’s ability to acquire, clean and transform source data will quickly amaze you, but those tasks are just the early steps in your analysis and visualization process. It wasn’t until 2008 that the pandas (http://pandas.pydata.org/) module was created by AQR Capital Management to provide “Pythonic” counterparts to the analytical foundations of languages like R, SAS or MATLAB, which is where the “real fun” begins.

While Python’s interpreter provides an interactive execution shell, aficionados recognized the need to extend this basic functionality and developed an even more dynamic and robust interactive environment—IPython—to fill the need. When coupled with the pandas module, any budding data analyst now has a mature and data-centric toolset available to drive their quest for knowledge.

Why R?

Unlike Python, R’s history is inexorably tied to its domain specific predecessors and cousins, as it is 100% focused and built for statistical data analysis and visualization. While it, too, can access and manipulate various file types and databases (and was also designed for flexibility and extensibility), R’s lisp- and S-like syntax and extreme focus on foundational analytics-oriented data types has kept it, mostly, in the hands of the “data crunchers”.

Base R makes it remarkably simple to run fundamental statistical analyses on your data and then generate informative and appealing core visualizations with just a few lines of code. More modern R libraries such as plyr and ggplot2 extend and enhance these base capabilities and are the foundations of many of mind- and eye-catching examples of modern data analysis and visualization you have no doubt come across on the internet.

Like Python, R also provides an interactive execution shell that has enough basic functionality for general needs. Yet, the desire for even more interactivity sparked the development of RStudio, which is a combination of IDE, data exploration tool, and iterative experimentation environment that exponentially enhances R’s default capabilities.

Why Both?

If all you have a hammer, everything starts looking like a nail. There are times when the flexibility of a general purpose programming language will come in very handy and that should definitely be a situation where you use Python. There are other times when three lines of R code will do something that may take thirty or more lines of Python code to accomplish. Since your ultimate goal is to provide insightful and accurate analyses as quickly and visually appealing as possible, knowing which tool to use for what job is a critical insight you must develop to be as effective and efficient as possible.

We would be a bit dishonest, though, if we did not concede that there are some things that Python can do (easily or at all) that R cannot, and vice-versa. We’ll touch upon some of these in the use-cases throughout the book, but many of the—ah—“learning opportunities” will only come from performing your own analyses, getting frustrated (which is the polite way of saying “stuck”) and only finding resolution by jumping to another tool to “get stuff done”. This situation comes up frequently enough that there is even a “rJython” package for R that lets you call Python code from R and “rpy” and “rpy2” modules for Python that let you call R code from Python.

By having both tools in your toolbox, you should be able to tackle most, if not all, of the tasks that come your way. Both R and Python have vibrant communities that are eager to provide assistance and even help in the development of new functions or modules to fit emerging needs.

Jumpstarting Your Python Analytics With Canopy

It *is* possible to setup an effective and efficient installation of Python, IPython and pandas from the links we’ve provided, especially if you are already familiar or proficient with Python. However, even the most stalwart Python aficionado can find it challenging to manage dependencies and updates for the numerous necessary components. This is especially true if you have to manage analytics processes across multiple operating systems and environments.

To facilitate both ease of installation and maintenance, we highly recommend using the freely available Enthought Canopy Python data analysis environment (<https://www.enthought.com/products/canopy/>). Canopy works on Linux, Microsoft Windows and Mac OS X; has a built-in Python integrated development environment (IDE); incorporates a meta- package manager that will help you keep current with changes in every dependent package and module and also comes with an IPython console. For those working in organizations that shy away from open source solutions, Enthought also offers commercially supported options for Canopy.

Given that there is a comprehensive installation, setup and update guide available (http://docs.enthought.com/canopy/quick-start.html), we will not go over step-by-step instructions on how to install Canopy for each platform, but we strongly recommend reviewing the documentation before attempting any of the Python examples in the book. Once the base installation is complete getting started should be as straightforward as opening up the Canopy application, which will display the welcome screen (Figure 2.1).

Figure 2.1 Canopy Welcome Screen [793725c02f001.png]

One of first steps you should perform is to instruct Canopy to display all images *inline* within the IPython console. This is an optional step, but it will help keep all output self-contained within the Canopy environment. You can change this setting once you have an open editor by going into the *Preferences* dialog, finding the *Python* tab and selecting the “*Inline (SVG)*” option for the “*PyLab backend*:” preference.

Figure 2.2 Canopy IDE With Preferences Open [793725c02f002.png]

To validate that your environment is setup properly, run the following code in the IPython console area in the editor:

import pandas as pd

import numpy as np

np.random.seed(1492)

test\_df = pd.DataFrame({ "var1": np.random.randn(5000) })

test\_df.hist()

and verify it produces the output seen in Figure 2.3. If it does, you have the basic environment ready to start working through the data analysis examples. If the bar chart is not displayed, you may need to check either your installation or verify you have the proper graphics display options mentioned earlier.

Figure 2.3 Test IPython Console Output [793725c02f003.png]

Once everything is working properly, you should carve out ten minutes to read through “*Learn Python in 10 minutes*” (<http://www.stavros.io/tutorials/python/>) by Stavros Korokithakis if you are not familiar with Python and then ten additional minutes to go through the “*10 Minutes to Pandas*” tutorial (<http://pandas.pydata.org/pandas-docs/dev/10min.html>) to learn a bit more about Pandas.

Understanding The Python Data Analysis And Visualization Ecosystem

While there are scores of modules available for Python, a few stand out when it comes to crunching data. Here are some that you will find yourself using in nearly every project:

* *NumPy* (<http://www.numpy.org/>) – a library providing foundational capabilities for creating multi-dimensional containers of generic data, performing a wide range of operations on data and generating random numbers. It also implements the capability to “broadcast” operations to Python objects, which can make for succinct and highly efficient code;
* *SciPy library* (<http://www.scipy.org/scipylib/index.html>) – built on top of NumPy, this library makes quick work of array-oriented operations and provides a facility to expand NumPy’s “broadcast” operations to other types of data elements in Python; it also provides additional statistical operations;
* *Matplotlib* (<http://matplotlib.org/>) – the de-facto way to turn your data into production-quality images in Python;
* *Pandas* – a library providing high-performance, easy-to-use data structures and data analysis tools; pandas introduces the Data.Frame type into the Python namespace which we will discuss in more detail later in this chapter; while this may cause some die-hard Python folks to cringe, pandas—in essence—makes Python more like R and should make it easier for you to jump between languages.

These modules, combined with IPython, are sometimes referred to the core components of the “*SciPy stack*” (which is confusing, since it contains the “*SciPy library*”), and you can read more about the stack at <http://www.scipy.org/>.

As you make your way through this ecosystem, you will notice the following code pattern emerge:

import numpy as np

import scipy as sp

import matplotlib as mpl

import matplotlib.pyplot as plt

import pandas as pd

The import statement brings the namespace and functionality of a particular library/module into the current Python working session and the “as” component of the statement provides an abbreviated reference for the functions, objects and variables in the module. Since you’ll be using many of the components of each of the modules in the *SciPy stack* on a regular basis you will save time and typing if you use a base template file.

type="note"

Python “Gotchas”

There are two features of Python that are liable to both frustrate and perhaps become problematic for new users. The first “gotcha” is whitespace. Spaces are significant in Python code. There are no “{}” braces or “begin/end” pairs to signify a block of code. You must use consistent indentation to identify groups of statements that will execute together. Inconsistency will result either in error messages from the interpreter or your code failing to work as expected. Most modern text editors or IDE can be configured to take care of this for you.

The second “gotcha” is the lack of a requirement to declare variables before using them. Initializing a variable named “breaches” to some value then inadvertently referring to it later as “breached” may not throw an error in the interpreter, but will most assuredly generate unexpected output.

You will, of course, use other packages for connecting to databases, reading from files and performing other functions and you can burn countless hours perusing all the nifty modules at the Python Package Index (PyPI), https://pypi.python.org/pypi, but the ones associated with the *SciPy stack* will become familiar and regular companions on your data science journey.

Setting Up Your R Environment

To build your R/RStudio environment you will need to download and install R (<http://cran.rstudio.com/>) then do the same for RStudio (<http://www.rstudio.com/ide/download/>). Both links provide full installation details for Linux, Windows and Mac OS X systems so we won’t delve into the minutiae in this section. You will, however, need to make a choice when you install RStudio as it comes in two flavors: Desktop and Server. Both provide the same core features:

* built-in IDE
* data structure and workspace exploration tools
* quick access to the R console
* R help viewer
* graphics panel viewer
* file system explorer
* package manager
* integration with version control systems

The primary difference is that one runs as a standalone, single-user application (RStudio Desktop) and the other (RStudio Server) is installed on a server, accessed via browser and enables multiple users to take advantage of the compute infrastructure. If you are not familiar with R or RStudio, begin by downloading and installing RStudio Desktop (all examples in this book involving RStudio assume you are working in the Desktop version).

For those who are limited to commercially supported tools, Revolution Analytics (<http://www.revolutionanalytics.com/support/>) provides commercial offerings and technical support for R.

Once everything is installed, open RStudio and verify you see the default workspace, which should look similar to Figure 2.4.

Figure 2.4 RStudio Default Workspace [793725c02f004.png]

If all is working correctly, you should take some time to walk through “*A (very) short introduction to R*” by Paul Torfs and Claudia Brauer (<http://cran.r-project.org/doc/contrib/Torfs%2BBrauer-Short-R-Intro.pdf>). It will run through just enough of the basics of the R language and RStudio environment to make you dangerous.

While you can use the built-in package manager with RStudio to install packages, you will eventually come to the realization that using the console method much more convenient. To get familiar with this process right away, you should install the “ggplot2” package, which is the primary graphics library we’ll be using in our examples and as straightforward as entering the following into the RStudio console pane:

**> install.packages(“ggplot2”)**

Installing package(s) into ‘/Library/Frameworks/R.framework/Versions/2.15/Resources/library’

(as ‘lib’ is unspecified)

trying URL 'http://cran.mirrors.hoobly.com/bin/macosx/leopard/contrib/2.15/ggplot2\_0.9.3.1.tgz'

Content type 'application/x-gzip' length 2659920 bytes (2.5 Mb)

opened URL

==================================================

downloaded 2.5 Mb

The downloaded binary packages are in

/var/folders/qg/vmtfcv1j7vjfq\_p5zw86mk7mxkhymk/T//RtmpiZ5FD3/downloaded\_packages

Run the following code to verify it has been installed correctly and to ensure your R/RStudio environment is functional:

library(ggplot2)

set.seed(1492)

test.df = data.frame(var1=rnorm(5000))

ggplot(data=test.df) + geom\_histogram(aes(x=var1))

If there are no errors and you see the bar chart in Figure 2.5 your environment has been setup correctly.

Figure 2.5 Test R/RStudio Output [793725c02f005.eps]

Like Python, R has a vast repository of useful modules that can simplify many tasks. You will be introduced to a few of them in the coming chapters but should also peruse the Comprehensive R Archive Network (CRAN) (<http://cran.r-project.org/web/packages/>) to see the breadth and depth covered by a host of contributors.

Organizing Analyses

As you prepare to jump into into data analysis projects, it’s good idea to setup an area where we you organize input data, analysis scripts, output (visualizations, reports and/or data) and any supporting documentation. For the purposes of the examples in this book, we’ll be using the following directory structure:

/book/ch03

|-*R*

|-data

|-docs

|-output

|-python

|-support

|-tmp

Like most elements of programming, there is no one, true way to setup this structure, but you should strive to find one that works for you and stick with it. A great way to do that is to take a lesson from modern web framework builders and use a simple setup script that builds the structure for you:

Sample analysis preparation script (Bourne shell script)

#!/bin/sh

#

# prep: prep analytics directory structure

#

# usage: prep DIRNAME

#

DIR=$1

if [ ! -d "${DIR}" ]; then

mkdir -p ${DIR}/*R* \

${DIR}/data \

${DIR}/docs \

${DIR}/output \

${DIR}/python \

${DIR}/support \

${DIR}/tmp

> ${DIR}/readme.md

ls -lR ${DIR}

else

echo "Directory "${DIR}" already exists"

fi

Sample analysis preparation script (Windows shell script)

SET PDIR=%1

IF EXIST %%PDIR GOTO HAVEPDIR

MKDIR %%PDIR

MKDIR %%PDIR\R

MKDIR %%PDIR\data

MKDIR %%PDIR\docs

MKDIR %%PDIR\output

MKDIR %%PDIR\python

MKDIR %%PDIR\support

MKDIR %%PDIR\tmp

<NUL (SET/P Z=) >%%PDIR\readme.md

DIR %%PDIR

:HAVEPDIR

ECHO "Directory exists"

You now only need to type “prep NAME” whenver you want to start a new project (so, for this project, “prep ch03”). As you develop your own styles and patterns, you can expand this script to include the generation of various templates and even initialization of source code repositories. Once the structure is in place, it’s time to retrieve, explore and analyze some data.