DataCamp
Learn data analysis for free, interactively

DATA SCIENCE WARS

VS.

R and Python are waging war:
while both programming languages are gaining prominence in the data analytics community, they are fighting to become data scientists' language of choice.
Which side are you taking?

Introducing The Opponents

Current Version

3.1.3 March 2015	3.4.3 / 2.7.9 February 2015/ December 2014

History

Creators	Creator
Ross Ihaka and Robert Gentleman	Guido Van Rossum
Release Year	Release Year
1995	1991
Must Knows	Must Knows
1. R is an implementation of S	1. Python was inspired by C, Modula-3,

programming language (Bell Labs).

2. R's design and evolution is handled by the R-core group and R foundation.

3. R's software environment was written primarily in C, Fortran and R.

and particularly ABC.

2. Python gets its name from the "Monty Python's Flying Circus" comedy series.

3. Python Software Foundation (PSF) takes care of Python's advances.

Purpose

R focuses on better, user friendly data analysis, statistics and graphical models.

Python emphasizes productivity and code readability.

Used By?

R has been used primarily in academics and research. However, R is rapidly expanding into the enterprise market.

"The closer you are to statistics, research and data science, the more you might prefer R."

Python is used by programmers that want to delve into data analysis or apply statistical techniques, and by developers that turn to data science.

"The closer you are to working in an engineering environment, the more you might prefer Python."

Community

Huge community with support coming in the form of:
- Mailing lists
- User-contributed documentation
- Active Stackoverflow members

More adoption from researchers, data scientists, statisticians, quants.

Overall good support for general purpose coding. Python support is found at:
- Stackoverflow
- Mailing lists
- User-contributed code and documentation

More adoption from developers and programmers.

Usability

Statistical models can be written with only a few lines.

There are R stylesheets but not everyone uses them.

The same piece of functionality can be written in several ways in R.

Coding and debugging is easier to do in Python, mainly because of the "nice" syntax.

The indentation of the code affects its meaning.

Any piece of functionality is always written the same way in Python.

Flexibility

It is easy to use complex formulas in R. All kinds of statistical tests and models are readily available and easily used.

Python is flexible for doing something novel that has never been done before. Developers can also use it for scripting a website or other applications.

Ease of Learning

R has a steep learning curve at start. Once you know the basics, you can easily learn advanced stuff.

R is not hard for experienced programmers.

Check out DataCamp's interactive exercises and tutorials.

Python's focus on readability and simplicity makes that its learning curve is relatively low and gradual.

Python is considered a good language for starting programmers.

Try using the book "Learn Python The Hard Way" and its accompanying site with videos and exercises.

Code Repositories

CRAN stands for the Comprehensive R Archive Network: it is a huge repository of R packages to which users can easily contribute.

Packages are collections of R functions, data, and compiled code. They can be installed in R with one line.

PyPi is the Python Package Index: it is a repository of Python software, consisting of libraries. Users can contribute to Pypi, but it is a bit complicated in practice.

Watch out with dependencies and installing Python libraries!

"I don't see Python [...] building up a huge code repository comparable to CRAN. [R has] a gigantic head start, [and] [...] statistics simply is not Python's central mission;"
- Norm Matloff, professor of computer science

Miscellaneous

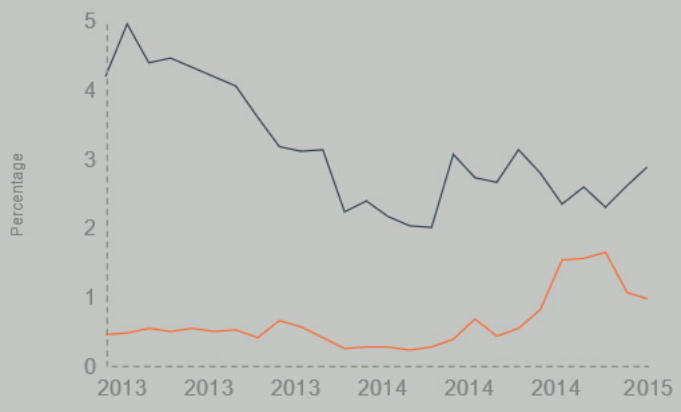
Use the rPython package to run Python code from R. Pass or get data from Python, call Python functions or methods.

Use the RPy2 library to run R code from within Python. It provides a low-level interface from Python to R.

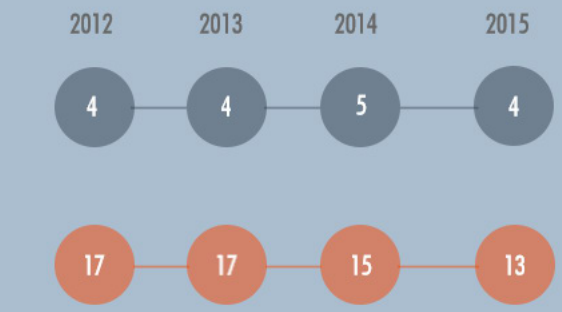
R and Python: The Numbers

Popularity Rankings

R and Pythons popularity between 2013 and February 2015 (Tiobe Index)



Redmonk ranking, comparing the relative performance of programming languages on GitHub and Stack Overflow (September 2012 and January 2013, 2014, 2015)



Jobs And Salary?

2014 Dice Tech Salary Survey:
Average Salary For High Paying Skills and Experience

$ 115,531

$94,139

The Data Analysis Battlefield

Usage

R is mainly used when the data analysis tasks require standalone computing or analysis on individual servers.

Python is generally used when the data analysis tasks need to be integrated with web apps or if statistics code needs to be incorporated into a production database.

Task

For exploratory work, R is easier for beginners. Statistical models can be written with a few lines of code.

As a full-fledged programming language, Python is a good tool to implement algorithms for production use.

Data Handling Capabilities

R is handy for data analysis because of the huge number of packages, readily usable tests and the advantage of using formulas.

R is usable for basic data analysis without the installation of packages. Big datasets require the use of packages such as data.table and dplyr, though.

The infancy of Python packages for data analysis was an issue in the past, but this has improved a lot!

You need to use NumPy and pandas (amongst others) to make Python usable for data analysis.

Getting Started

IDE

RStudio

Popular Packages

- ✓ dplyr, plyr and data.table to easily manipulate data.
- ✓ stringr to manipulate strings.
- ✓ zoo to work with regular and irregular time series
- ✓ ggvis, lattice and ggplot2 to visualize data.
- ✓ caret for machine learning.

Tip: check out DataCamp's online

IDE

There are many Python IDEs to chose from. However, Spyder and IPython Notebook are most popular.

Tip: also look up Rodeo, the "data science IDE for Python"

Popular Libraries

- ✓ pandas to easily manipulate data.
- ✓ SciPy /NumPy for scientific computing.
- ✓ sckikit-learn to use machine learning methods.
- ✓ matplotlib to make graphics.
- ✓ statsmodels to explore data, estimate

interactive courses and tutorials!

tests and unit tests.

"R is currently head-and-shoulders above Python for data analysis, but I remain convinced that Python CAN catch up, easily and quickly."
- Jan Galkowski, computational engineer

Support

There's a lot of support out there for data analysis with R:

- ✓ Stackoverflow
- ✓ Rdocumentation , the R documentation aggregator
- ✓ R-help mailing list

Support for data analysis issues can be found at:

- ✓ Stackoverflow
- ✓ Mailing lists:

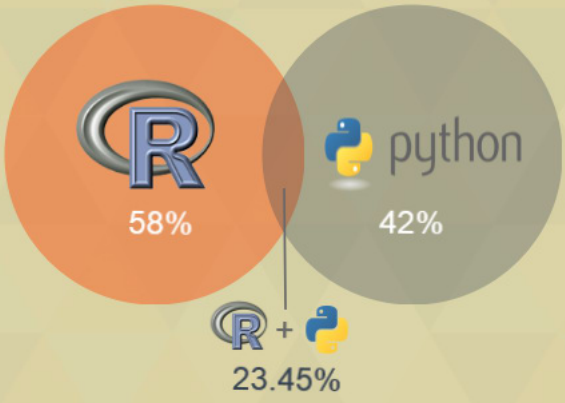
pydata	Questions related to Python for data analysis and pandas
pystatsmodels	Statsmodels or pandas questions
numpy-discussion	Numpy questions
sci-py user	General SciPy or scientific questions

R And Python: The Quantified Battlefield

General

Languages for data analysis used in 2014 (KDnuggets polls)

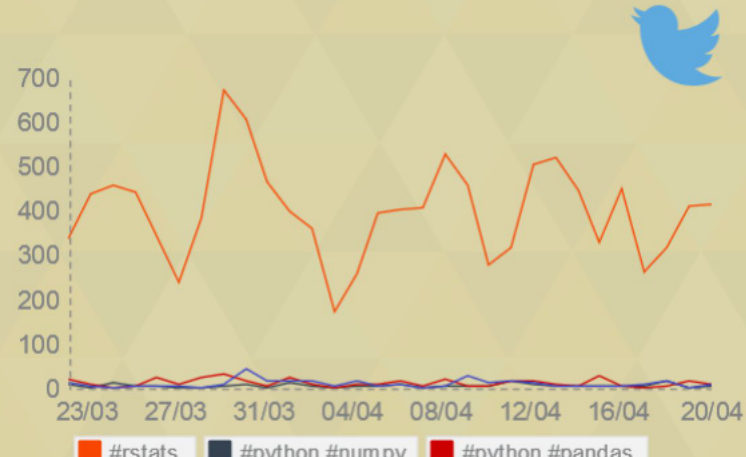
Analysis of R and Python used together in 2014 (KDnuggets polls)



Community?

Stack Overflow Questions tagged "R" and/or "Python", "Pandas" between 2008 and April 15, 2015

Twitter activities between March 12 and April 10, 2015



R Pandas NumPy

#numpy

Jobs and Salary?

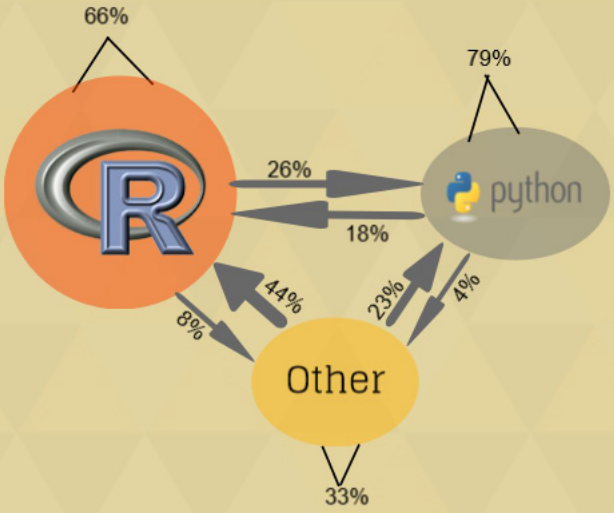
O'Reilly 2014 Data Science Salary Survey

Average Annual Salaries In The Range Of:

Python = US$110, 000 to US$125,000 = R

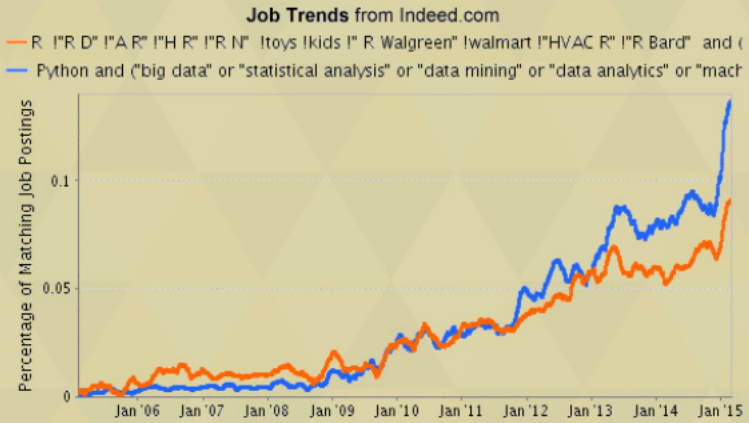
Switching Between R and Python?

Number of people switching between R and Python in 2013 *



*Percentages on the arrows are relative to the base

R and Python job trends



"My current strategy is to leverage the best of both worlds — do early stage data analysis in R, then switch to Python when it's time to get serious, be a team player, and ship some real code and data products."

• • •

"I use R to conduct statistical tests, graph data, and inspect large data sets. If I actually have to write an algorithm, I prefer Python..."

• • •

"I'd rather do math in a general-purpose language than try to do general-purpose programming in a math language."

#3 The Last Stand: Pros And Cons

Graphical Capabilities

A picture says more than a thousand words

Visualized data can be understood more efficiently and effectively than the raw numbers alone.

R + visualization = perfect match

ggplot2 To make pretty graphs, including the

+ IPython Notebook

Bundle your analysis in one file

The IPython Notebook makes it easier to work with Python and data.

Simplify your workflow when working with data in Python

It's a combination of:

create layered, customizable plots

lattice To easily display multivariate relationships

rCharts To create, customize and publish interactive javascript visualizations from R

googleVis To use Google Chart tools to visualize data in R

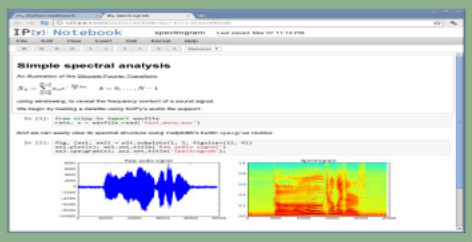
ggvis To implement interactive grammar of graphics, while rendering in a web browser

e.g.: Visualizing Facebook friends with R

Interactive python exploration, prewritten programs, text, and equations for documentation in one environment

Share notebooks with colleagues without having them install anything.

The IPython notebook drastically reduces the overhead of organizing code, output, and notes files, which allows to spend more time doing real work.



The R Ecosystem

The R Project

Rich ecosystem of cutting-edge interface packages available to communicate between open-source languages.

This allows you to string your workflow together, which is especially useful for data analysis.

Packages are available at:

Cran "Task Views" page lists a wide range of tasks for which R packages are available

Bioconductor Open source software for bioinformatics

GitHub web-based Git repository hosting service

Rdocumentation

GitHub NEW!

Bioconductor NEW!

CRAN

Search through all these sources easily with Rdocumentation, the first R documentation aggregator

The R User Community

✓ Meetup groups

Some are sponsored by companies of the R community

REVOLUTION ANALYTICS

MANGO SOLUTIONS

Python, A General Purpose Language

Readability and Learning Curve

Just like everyday English

Python is easy and intuitive, and its emphasis on readability only magnifies these characteristics.

e.g. `print("Hello World!")`

Syntactically clear and elegant code, easily interpretable and very easy to type.

This explains why.

✓ Python's learning curve is relatively flat

✓ So many programmers are familiar with it

Also, the speed at which you can write a program is also positively impacted:

Less time coding, more time playing

The Python Testing Framework

Guarantee your code is reusable and dependable

A built-in, low barrier-to-entry testing framework that encourages good test coverage.

Python Testing Tools Taxonomy, including

UnitTest First unit test framework of the Python standard library

Nose Extends UnitTest: used in many packages such as pandas

DocTest Easy generation of tests based on output

✓ Blogs & Social Media

R-bloggers #rstats

from the standard Python interpreter shell

Pytest — To write small tests, while supporting complex functional testing

testing-in-python (TIP) mailing list

R, Lingua Franca of Statistics

Developed by statisticians, for statisticians

Statisticians communicate ideas and methods for statistical analysis through R code and packages.

Statisticians, engineers and scientists without computer programming skills find it easy to use.

Increasing industry adoption...

R is used in finance, pharmaceuticals, media and marketing; In this last area, R's on the rise as a business analytics tool.

"The number one value to businesses in using R is access to talent"

Google  

... And widespread use in academia

R is experiencing a rapid growth, solidifying its position in third place as software used in scholarly articles, right after SAS and SAP.

Python, A Multi-Purpose Language

Ready To Work!

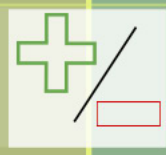
As a common, easy-to-understand language that is known by many programmers, Python also brings people with different backgrounds together.

For example,

Some organizations that didn't want to hire or had difficulties to hire new data scientists (re)trained their existing employees to use Python instead.

This means that Python is a production ready language: it has the capacity to be a single tool that integrates with every part of your workflow!

R Is Slow



R is slow, on purpose

R was designed to make data analysis and statistics easier to do, not to make life easier for your computer.

R has an incomplete informal definition; It is mostly defined in terms of how its implementation works.

Beyond design and implementation, a lot of R code is slow simply because it's poorly written.

Packages to improve R's performance:

pqR	A new version of the R interpreter
renjin, FastR	Original R rewritten in Java
Riposte	A fast interpreter and JIT for R
RevoScaleR	Commercial tool to handle big datasets
Foreach	Commercial tool that acilitates parallel

Python And Visualizations

"Visualizations are important criteria in choosing data analysis software"

Python has some nice visualization libraries:

Seaborn	Library based on matplotlib
Bokeh	Interactive visualization library
Pygal	To create dynamic svg charts

But there are a lot of options to choose from; Maybe too many.

Moreover, in comparison to R

"Visualizations in Python are usually more convoluted, and the results are not nearly as pleasing to the eye or as informative."

programming

R's Steep Learning Curve

"The worst thing about R is that ... it was developed by statisticians."

R's learning curve is nontrivial:

- Even though anybody can get results using GUIs, none is comprehensive enough to totally avoid programming.
- Finding packages can be time consuming

Using the right tools

Good resources can help you to overcome this steep learning curve:

DataCamp 's interactive exercises and tutorials

Rdocumentation to search for packages

Python Is Immature ("It's a challenger!")

A more limited way to think about data analysis

At the moment, there are no module replacements for the 100s of essential R packages

Python's catching up, but will this make people give up R?

- IPython's R extension allows you to cleanly use R in the IPython notebook.
- The current landscape of conventions and resources plays a huge role:

Matlab	Commonly used to publish open research code
Python	Used in mathematics
R	Used in statistics

Mlabwrap offers a bridge from Python to Matlab, but there are some drawbacks:

- You need to work with two languages
- You need a Matlab license

Shared Positive Points

Open-Source

R and Python are free to download for everyone, in comparison to other statistical software such as SAS and SPSS, which are commercial tools.

Advanced Tools

Many new developments in statistics appear first in the open source packages of R and, to lesser extent, Python, before making their way to commercial platforms.

Online Communities

While commercial softwares offer (paid) customer support, R and Python dispose of online communities that offer support to their respective users.

Paycheck

According to the O'Reilly 2013 Data Science Salary Survey, data scientists that use primarily open-source tools earned a higher median salary (US$130,000) than those using proprietary tools (US$90,000)

#4 And The Winner is...

It's a tie!
It's up to you, the data scientist,
to pick the language that best fits your needs.
The following questions can guide you in your decision.

1. What problems do you want to solve?

2. What are the net costs for learning a language?*

   * it will cost time to learn a new system that is better aligned for the problem you want to solve, but staying with the system you know may not be made for that kind of problem.

3. What are the commonly used tool(s) in your field?

4. What are the other available tools in your field and how do these relate to the commonly used tool(s)?

Sources :
http://pgbovine.net/ipython-notebook-first-impressions.htm
http://ipython.org/notebook.html
http://www.kdnuggets.com/2013/12/poll-results-r-leading-python-gaining.html
http://www.kdnuggets.com/polls/2014/languages-analytics-data-mining-data-science.html
http://matloff.wordpress.com/2014/05/21/r-beats-python-r-beats-julia-anyone-else-wanna-challenge-r
http://www.stat.washington.edu/~hoytak/blog/whypython.html
http://climateecology.wordpress.com
http://www.computerworld.com/article/2475559/big-data/is-python-really-supplanting-r-for-data-work.html
http://readwrite.com/2013/11/25/python-displacing-r-as-the-programming-language-for-data-science
http://www.kaggle.com/forums/t/5243/pros-and-cons-of-r-vs-python-sci-kit-learn
http://www.talyarkoni.org/blog/2012/06/08/r-the-master-troll-of-statistical-languages/
http://www.talyarkoni.org/blog/2013/11/18/the-homogenization-of-scientific-computing-or-why-python-is-steadily-eating-other-languages-lunch/
http://www.analyticsvidhya.com/blog/2014/03/sas-vs-vs-python-tool-learn/
http://slendermeans.org/language-wars.html
https://wiki.python.org/moin/OrganizationsUsingPython
http://blog.revolutionanalytics.com/2014/05/companies-using-r-in-2014.html
http://www.tiobe.com/index.php/content/paperinfo/tpci/index.html
https://github.com/hadley/r-python
http://www.stat.washington.edu/~hoytak/blog/whypython.html
http://www.thehelloworldprogram.com/python/
why-python-should-be-the-first-programming-language-you-learn/
http://dataconomy.com/python-displacing-r-in-data-science/
http://www.experfy.com/blog/python-data-science/
http://www.statmethods.net/about/learningcurve.html
http://www.ibm.com/developerworks/library/bd-learnr/
http://www.r-bloggers.com/faster-higher-stonger-a-guide-to-speeding-up-r-code-for-busy-people/
http://adv-r.had.co.nz/Performance.html
https://dynamicecology.wordpress.com/2014/01/14/r-isnt-just-r-anymore/
http://www.nytimes.com/2009/01/07/technology/business-computing/
07program.html?pagewanted=all&_r=0
http://blog.revolutionanalytics.com/2013/11/the-rise-of-r-as-the-language-of-analytics.html
http://www.revolutionanalytics.com/r-user-group-sponsorship-program
http://inside-bigdata.com/2013/12/09/data-science-wars-python-vs-r/

http://inside-bigdata.com/2013/12/09/data-science-wars-python-vs-r/
http://paulbutler.org/archives/visualizing-facebook-friends/
https://support.rstudio.com/hc/en-us/articles/201057987-Quick-list-of-useful-R-packages
http://blog.revolutionanalytics.com/2013/12/r-and-python.html
https://www.elance.com/q/blog/2012/01/modern-language-wars.html
http://r4stats.com/articles/popularity/
http://blog.revolutionanalytics.com/2014/02/r-salary-surveys.html
http://blog.revolutionanalytics.com/2014/01/in-data-scientist-survey-r-is-the-most-used-tool-other-than-databases.html
http://www.oreilly.com/data/free/stratasurvey.csp
http://redmonk.com/sogrady/2012/02/08/language-rankings-2-2012/