R for data science

...

A (very) gentle introduction

PyData Cyprus 17.05.2018

What is R

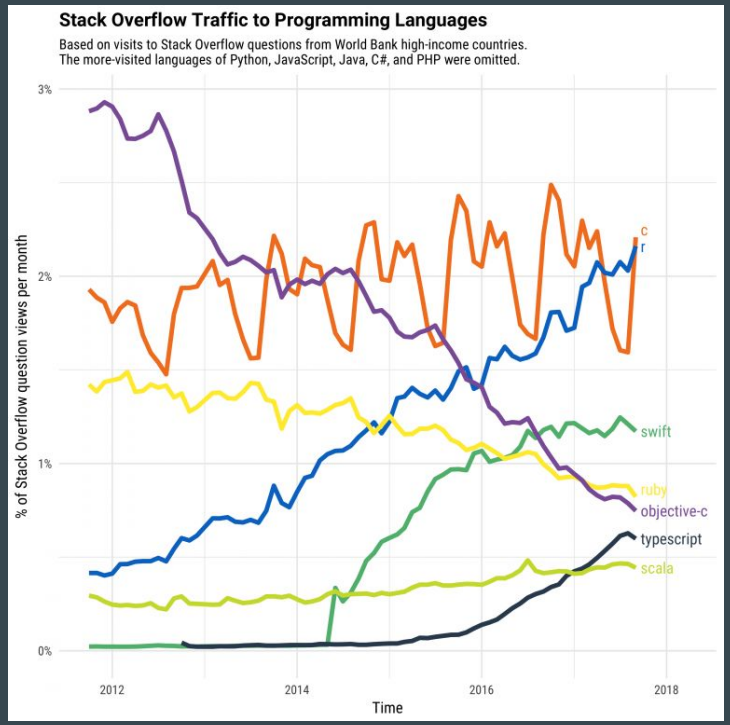
R is *a language and environment* for statistical computing and graphics.

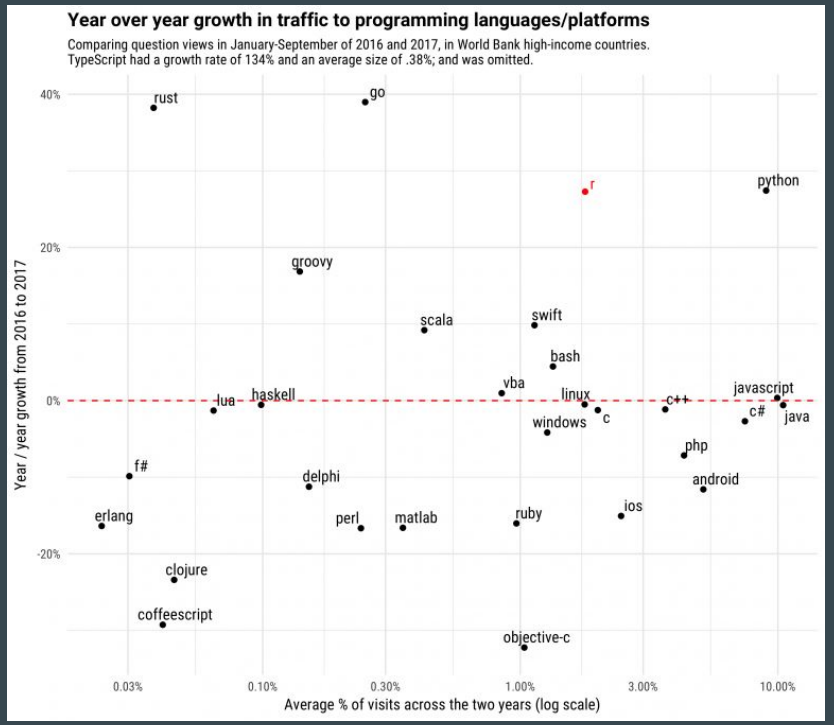
A GNU project

Originally the S language

Developed at Bell Laboratories (formerly AT&T, now Lucent Technologies) by John Chambers and colleagues.

R's rapid expansion





Source https://stackoverflow.blog/2017/10/10/impressive-growth-r/

R and machine learning

The CARET package - http://topepo.github.io/caret/index.html (Classification and Regression Training)

Text Mining with R - https://www.tidytextmining.com/

Neural Networks - https://beckmw.wordpress.com/tag/nnet/

Support Vector Machines - https://bit.ly/2L8gcvl

Deep learning using Keras and TensorFlow - https://keras.rstudio.com/, https://www.datacamp.com/community/tutorials/keras-r-deep-learning

Datasets in R

Built in datasets in R to play around with

https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/00Index.html

Examples:

- mtcars (good for visualisation practice)
- diamonds (good for visualisation practice)
- nycflights13 (good for data wrangling practice)
- gapminder (good for data wrangling practice)

[nycflights and gapminder must be loaded separately as a package - install.packages("nycflights13") and then library("nycflights13")

Getting started

1. Install R
   https://cran.r-project.org/ (the major depository of r packages)

2. Install R-Studio (the IDE for R)
   https://www.rstudio.com/

NOTE: You can operate R without R-Studio, but R-Studio needs a separate installation of R to work

Useful commands

When in need type ?+function_name

e.g. ?mean

Or ?package_name

e.g. ?ggplot2

Or ?dataset_name

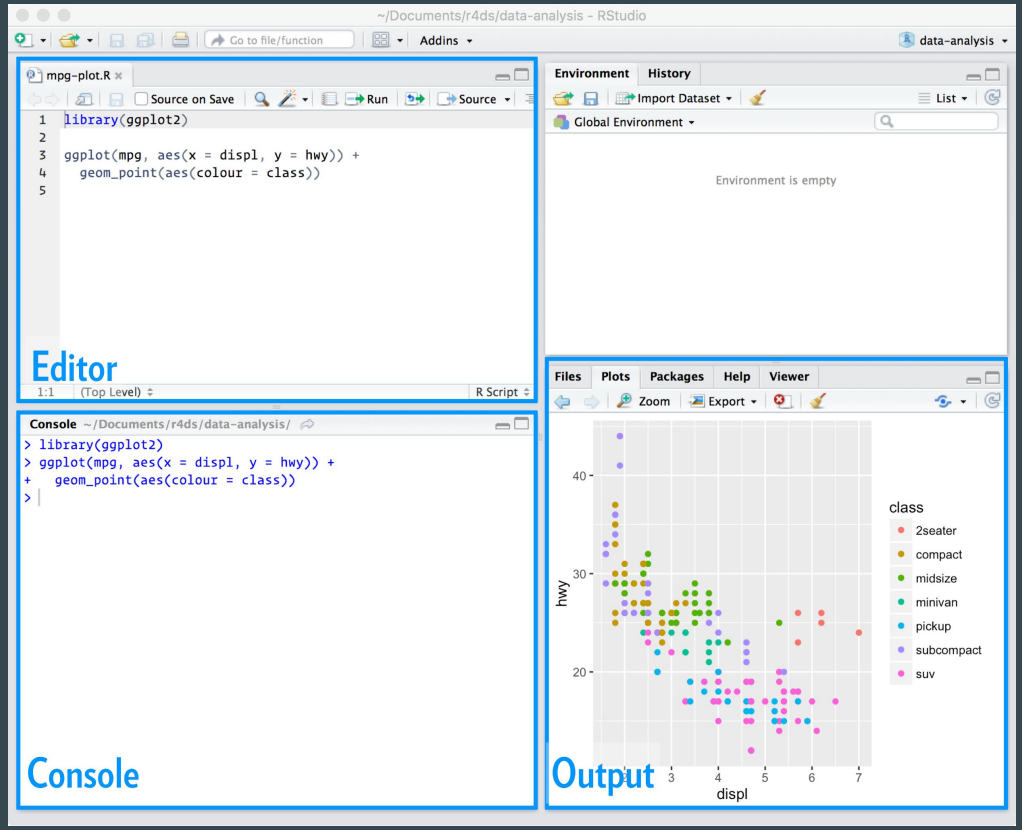
e.g. ?starwars

Workflow basics

http://r4ds.had.co.nz/workflow-basics.html

[lets get some exposure in R and R Studio]

R Studio



Cheatsheets - https://www.rstudio.com/resources/cheatsheets/

Genomics in R - https://www.bioconductor.org/ (A depository for genomics packages in R)

Courses in R

(Free) Introductory courses

https://www.datacamp.com/courses/free-introduction-to-r

https://www.datacamp.com/courses/intro-to-statistics-with-r-introduction

https://www.edx.org/course/introduction-to-r-for-data-science

https://www.udemy.com/r-basics/

Tutorials for Learning R - https://www.r-bloggers.com/how-to-learn-r-2/

Community

https://twitter.com/R4DScommunity

https://mobile.twitter.com/RLangTip

https://mobile.twitter.com/hashtag/rstats

https://mobile.twitter.com/Rbloggers

Resources

Books (free and online)

R for Data Science - http://r4ds.had.co.nz

Data visualization for Social Sciences - http://socviz.co/index.html

Introduction to Data Science - https://rafalab.github.io/dsbook/

Machine learning and statistical computing in R - http://r-statistics.co/

Cookbook for R - http://www.cookbook-r.com/ (code and stuff. Very good)

A Survival Guide to Data Science with R - https://togaware.com/onepager/

Text Mining with R - https://www.tidytextmining.com/

Blogs to follow

Variance explained - http://varianceexplained.org/ (good for data science using R and Python)

Mara Averick - https://maraaverick.rbind.io/ (good for visualisation)

R bloggers - https://www.r-bloggers.com/ (aggregating blogs writing about R- Extremely good)

Quick R - https://www.statmethods.net/index.html (tutorials etc. Extremely good)

Machine learning mastery - https://machinelearningmastery.com/start-here/ (good for machine learning)

Thank you!

Lefkios Paikousis
lefkiospaikousis@yahoo.co.uk
@lefkiospaik
linkedin.com/in/lefkios

PyData Cyprus 17.05.2018