ALGONQUIN COLLEGE

CST8390 BUSINESS INTELLIGENCE & DATA ANALYTICS

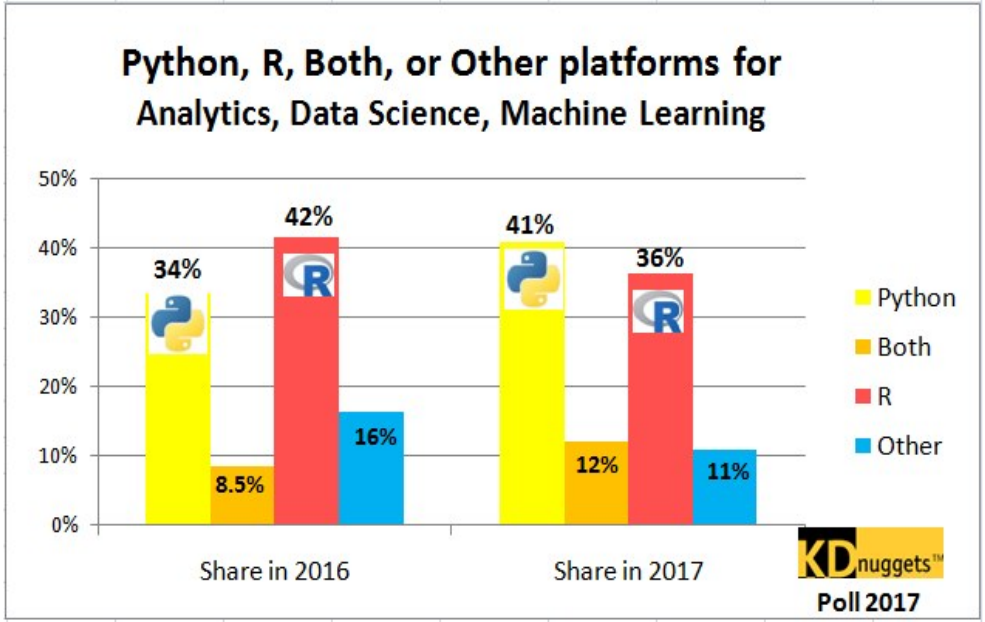
Data Science Tools

Professor : Dr. Anu Thomas
Email: thomasa@algonquincollege.com
Office: T314

Tools for Data Science

Data science tool	% of respondents using the tool
Python	76.3
R	59.2
SQL	53.6
Jupyter notebooks	40.3
TensorFlow	28.4
Amazon Web services	23.5
Unix shell / awk	23.3
Tableau	20.4
C/C++	19.2
NoSQL	19.2

Showing 1 to 10 of 49 entries Previous Next



Taken from https://blog.appliedai.com/data-science-tools/

Plug & Play Data Science Tools

- RapidMiner
- DataRobot
- BigML
- Google Cloud AutoML
- Paxata
- Trifacta
- MLBase
- Weka
- Driverless AI
- MS Azure ML Studio

Top Analytics/Data Science Tools

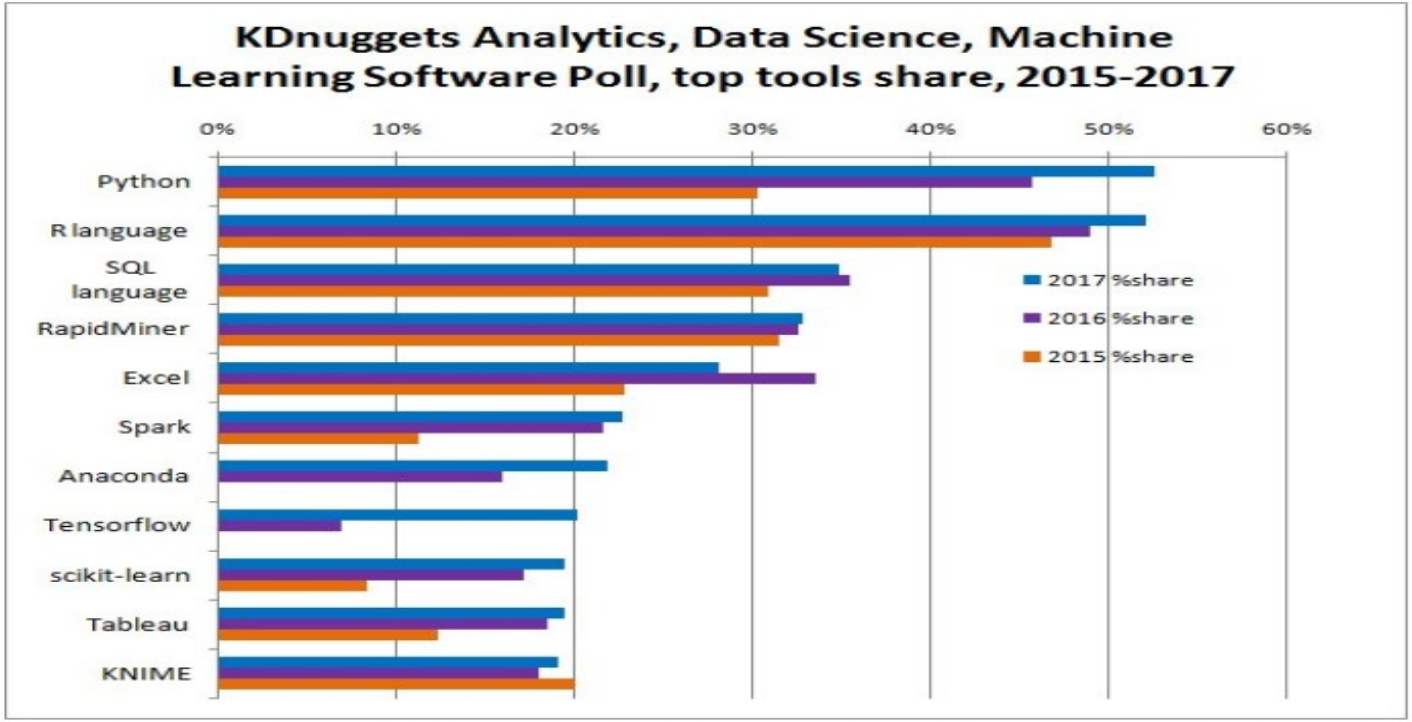


Fig 1: KDnuggets Analytics/Data Science 2017 Software Poll: top tools in 2017, and their usage in the 2015-6 polls

Taken from https://www.kdnuggets.com/2017/05/poll-analytics-data-science-machine-learning-software-leaders.html

Python - Introduction

- Python is an interpreted high-level programming language for general-purpose programming.
- Created by Guido van Rossum and first released in 1991
- Meant to be simple for non-programmers.
- Can be downloaded from https://www.python.org/

Python

- Python 3.7 is the latest version
- A lot of libraries available for various functionalities
- PyCharm – pretty good IDE for python (can be downloaded from https://www.jetbrains.com/pycharm/)

Python Libraries

Core Libraries	Visualization	Machine Learning	Statistics
NumPy	Matplotlib	Scikit-Learn (sklearn)	Statsmodels
SciPy	Seaborn		
Pandas	Bokeh		
	Plotly		

Python Libraries

- NumPy – fundamental package for scientific computing
  - a powerful homogeneous N-dimensional array object
  - sophisticated (broadcasting) functions
  - tools for integrating C/C++ and Fortran code
  - useful linear algebra, Fourier transform, and random number capabilities
- SciPy – library of software for engineering and science
  - SciPy contains modules for linear algebra, optimization, integration, and statistics.
  - The main functionality of SciPy library is built upon NumPy, and thus, its arrays make substantial use of NumPy.

Python Libraries

- Pandas - open source, BSD-licensed library providing high-performance, easy-to-use data structures and data analysis tools
  - Two main data structures – Series & DataFrames
  - Easily add and delete columns from a DataFrame
  - Convert data structures to DataFrame objects
  - Handle missing data, represented as NaNs
  - Group by functionality
- Scikit-Learn - a concise and consistent interface to common machine learning algorithms, making it simple to bring ML into production systems.
  - combines quality code and good documentation, ease of use and high performance, and is de-facto industry standard for machine learning with Python.

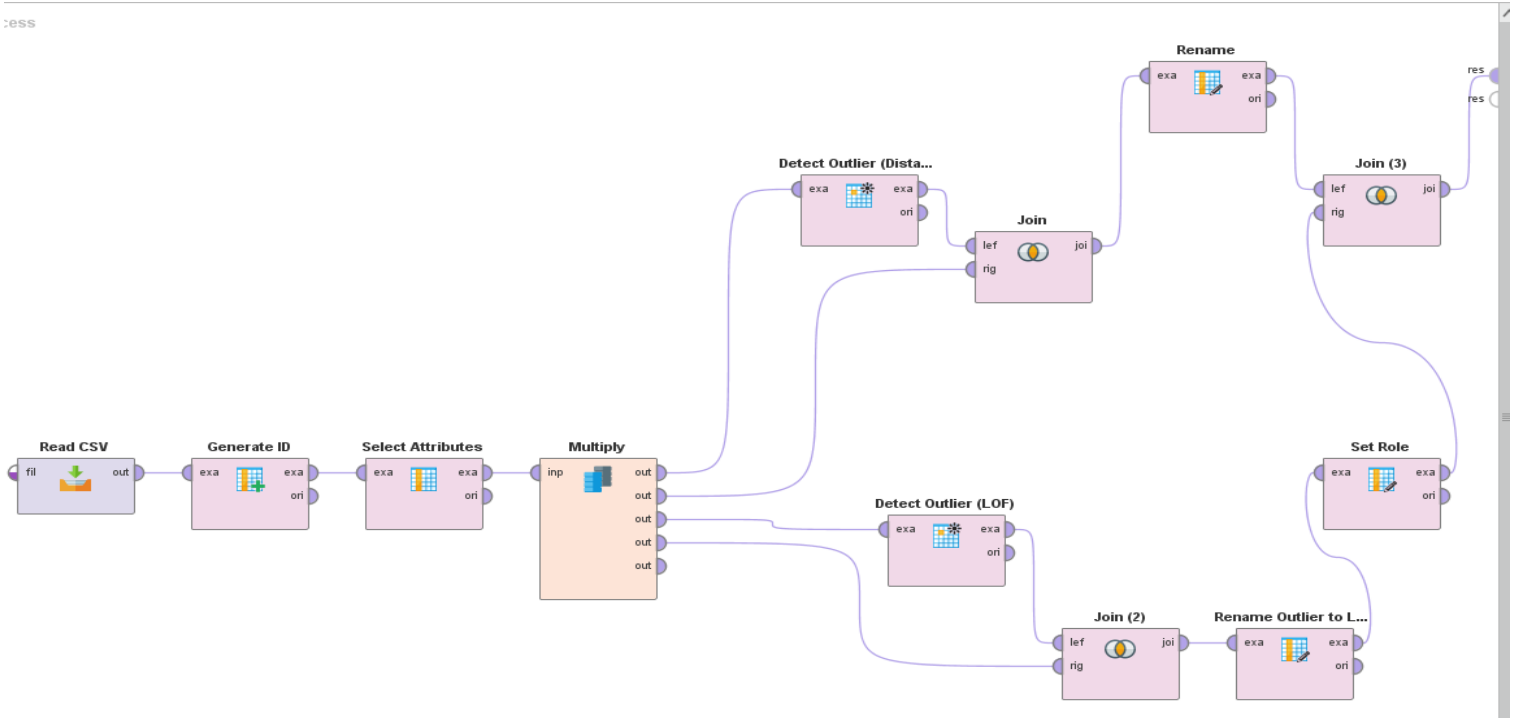
Python Demo

RapidMiner

- data science software platform that provides an integrated environment for
  - data preparation
  - machine learning
  - deep learning
  - text mining
  - predictive analytics

Can download from https://rapidminer.com/get-started/

RapidMiner Demo



References

- https://www.datascienceweekly.org/articles/what-tools-do-employers-want-data-scientists-to-know
- https://www.analyticsvidhya.com/blog/2018/05/19-data-science-tools-for-people-dont-understand-coding/
- https://www.kdnuggets.com/tag/data-science-tools
- https://www.python.org/
- https://www.jetbrains.com/pycharm/
- https://www.datascience.com/blog/top-python-libraries-for-data-science-in-2017