R

Instructor : YunGil Jun

Date: 2019. 11.

Written by YunGil Jun (Updated 2019. 11)

I. 데이터 분석을 위한 R

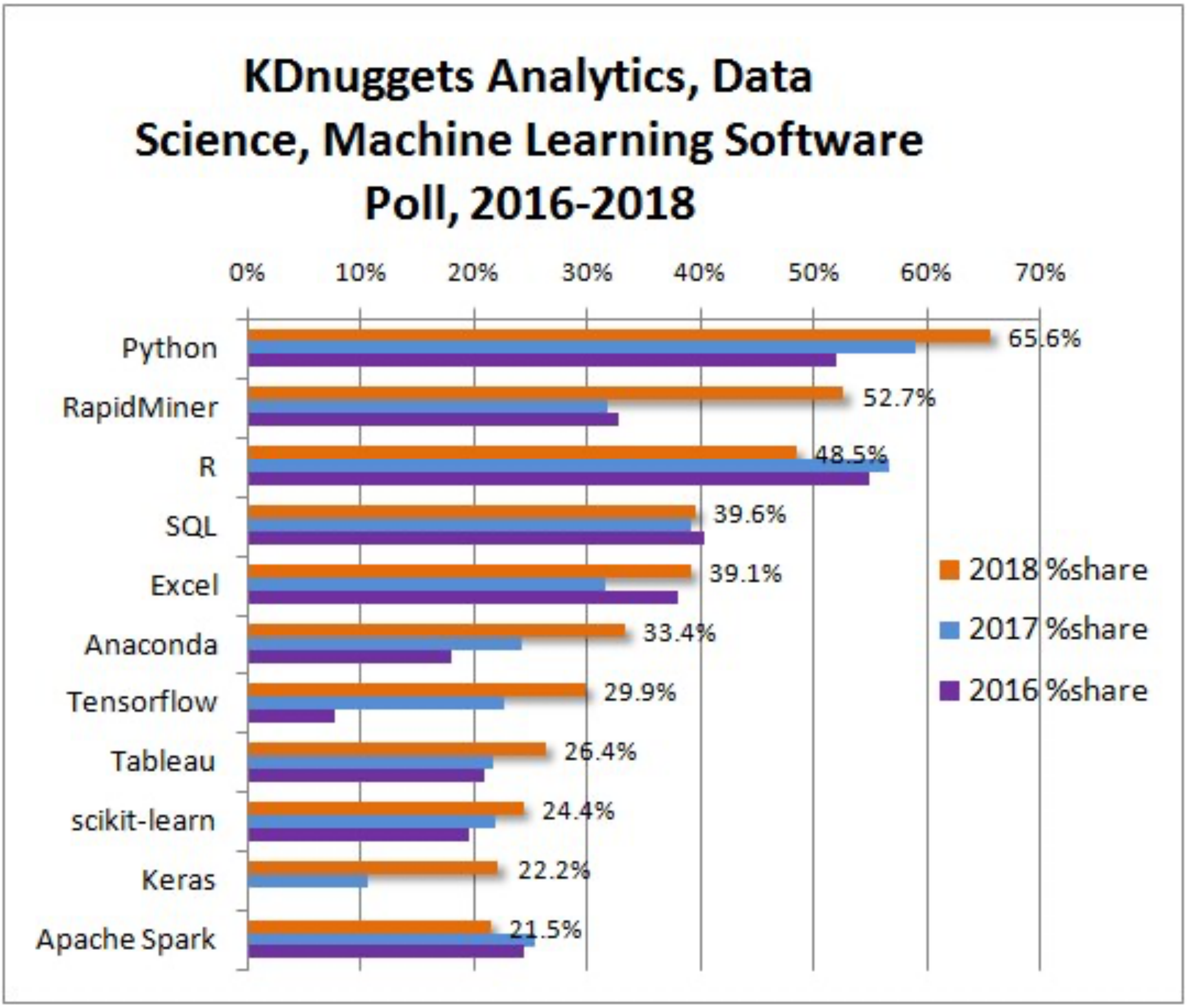
- 데이터 분석을 위한 오픈소스
- 데이터 분석 프로젝트 프로세스

1. What is R?

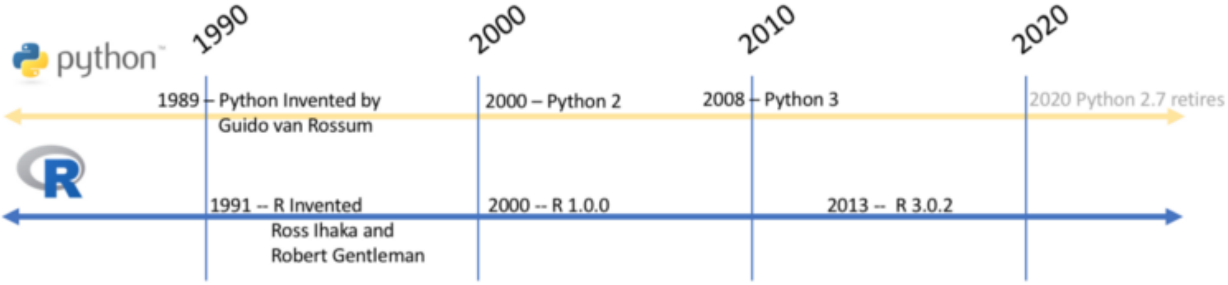
- **GPL(General Public License: GNU)** : 일반 공중 사용 허가서로 자유 소프트웨어 재단에서 만든 자유 소프트웨어 라이센스 (https://ko.wikipedia.org/wiki/GNU_일반_공중_사용권 (https://ko.wikipedia.org/wiki/GNU_일반_공중_사용권))
- #### R reference:
  - **공식사이트**: https://cran.r-project.org (https://cran.r-project.org)
  - **공식사이트 R 소개자료** : "An Introductin to R"(https://cran.r-project.org/doc/manuals/r-release/R-intro.html#Introduction-and-preliminaries (https://cran.r-project.org/doc/manuals/r-release/R-intro.html#Introduction-and-preliminaries))
  - "**R** is a language and environment for statistical computing and graphics"

2. Advantage of R

- **Comparison of Statistical Packages** (https://en.wikipedia.org/wiki/Comparison_of_statistical_packages (https://en.wikipedia.org/wiki/Comparison_of_statistical_packages))
- **Popularity for Data Scientists (R vs Python)**



- **History**



  - Python Invented(1989 Guido van Rossum) -> Python2(2000) -> Python3(2008)
  - S(Bell Labs) -> R Invented(1991 Ross Ihaka and Robert Gentleman) -> R 1.0.0(2000) -> R 3.0.2(2013)

comparision between R and Python for DataScientists (2019)

(written by author)

3. Why R?

- Open-source (Free)
- Lots of packages
- High quality packages with proper document (CRAN)
- easy to use with RStudio, R Markdown
- graphics capabilities
- community support

4. Environment in R

- 통합 개발 환경IDE(Integrated Development Environment)

> - RStudio : https://www.rstudio.com/products/RStudio/ (https://www.rstudio.com/products/RStudio/)
> - Visual Studio용 R : https://docs.microsoft.com/ko-kr/visualstudio/rtvs/installing-r-tools-for-visual-studio (https://docs.microsoft.com/ko-kr/visualstudio/rtvs/installing-r-tools-for-visual-studio)
> - jupyter notebook R kernel : https://irkernel.github.io/requirements/ (https://irkernel.github.io/requirements/)

5. R community

- #rstats hashtag : https://twitter.com/search?q=%23rstats (https://twitter.com/search?q=%23rstats)
- R-Ladies : https://rladies.org (https://rladies.org)
- Local R meetup groups : https://jumpingrivers.github.io/meetingsR/r-user-groups.html (https://jumpingrivers.github.io/meetingsR/r-user-groups.html)
- Rweekly : https://rweekly.org (https://rweekly.org)
- R-bloggers : https://www.r-bloggers.com (https://www.r-bloggers.com)
- DataCarpentry(http://www.datacarpentry.org (http://www.datacarpentry.org)) and Software Carpentry(https://software-carpentry.org (https://software-carpentry.org))
- R Conferences : https://jumpingrivers.github.io/meetingsR/events.html (https://jumpingrivers.github.io/meetingsR/events.html)
- Github : https://github.com/trending/developers/r?since=weekly (https://github.com/trending/developers/r?since=weekly)
- The R Consortium : https://www.r-consortium.org/projects (https://www.r-consortium.org/projects)

6. Etc

- Github, Jupyter Notebook, RMarkdown, etc
- Python(Deep learning), Julia(new) etc

참고) Computational methods for Analysis

1) Data Cleaning

- Importing data
- Joining multiple datasets
- Detecting missing values
- Detecting anomalies
- Imputing for missing values
- Data quality assurance

2) Exploratory Data Anlaysis(EDA)

- Ability to formulate relevant questions for investigation
- Identifying trends
- Identifying covariation between variables
- Communicating results effectively using visualizations(scatterplots, histograms, box and whisker, etc.)

3) Data Visualizations

- Including metrics relevant to your customer’s needs
- Creating useful features
- A logical layout (“F-pattern” for easy scanning)
- Creating an optimum refresh rate
- Generating reports or other automated actions

4) Analysis

```
(1) Knowledge based Analysis:
ex) Statistical Analysis

(2) Algorithmatic Analysis:
ex) Statistical Analysis, Machine Learning
```

(1) Statistical Analysis

- Emprical Study
- Theoritical Knowledge based Analysis

(2) Machine Learning

- Reason why you chose to use a specific machine learning model
- Splitting data into training/test sets (k-fold cross validation) to avoid overfitting
- Selecting the right evaluation metrics (AUC, $adj - R^2$, confusion matrix, etc.)
- Feature engineering and selection
- Hyperparameter tuning