

R for Data Science

(번역서) "R을 활용한 데이터과학" 발간에 맞춰

김설기

O'REILLY



R을 활용한 데이터 과학

데이터 불러오기, 정리하기, 변형하기, 시각화하기, 모델링하기

프로그래밍인사이트

R을 활용한 데이터 과학
데이터 불러오기, 정리하기, 변형하기, 시각화하기, 모델링하기

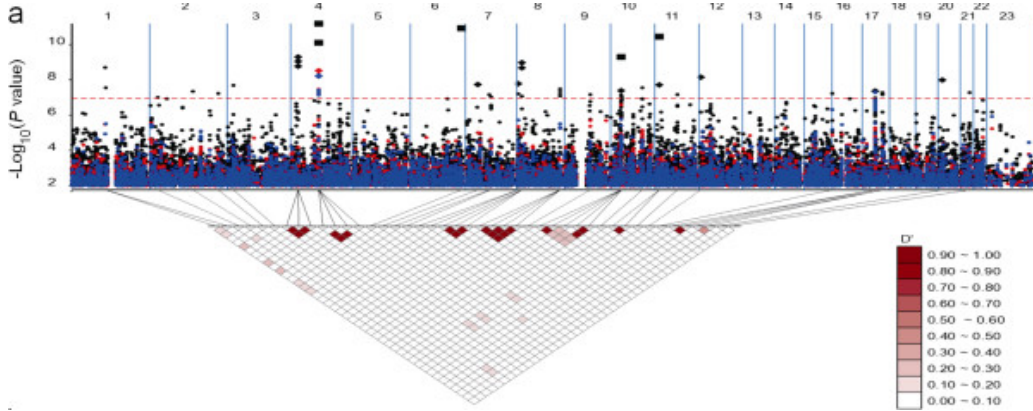
해설과 예제, 개념 그물망도 제공
강의기, 과제집, 실습용 데이터

인사이트
O'REILLY

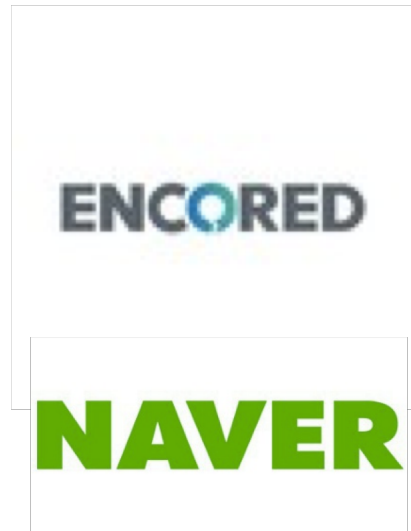
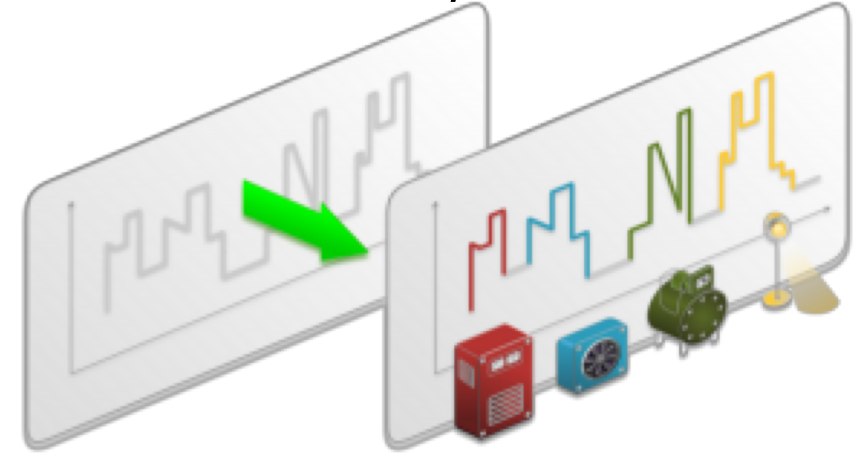
프로그래밍인사이트

해설과 예제, 개념 그물망도 제공
강의기, 과제집, 실습용 데이터

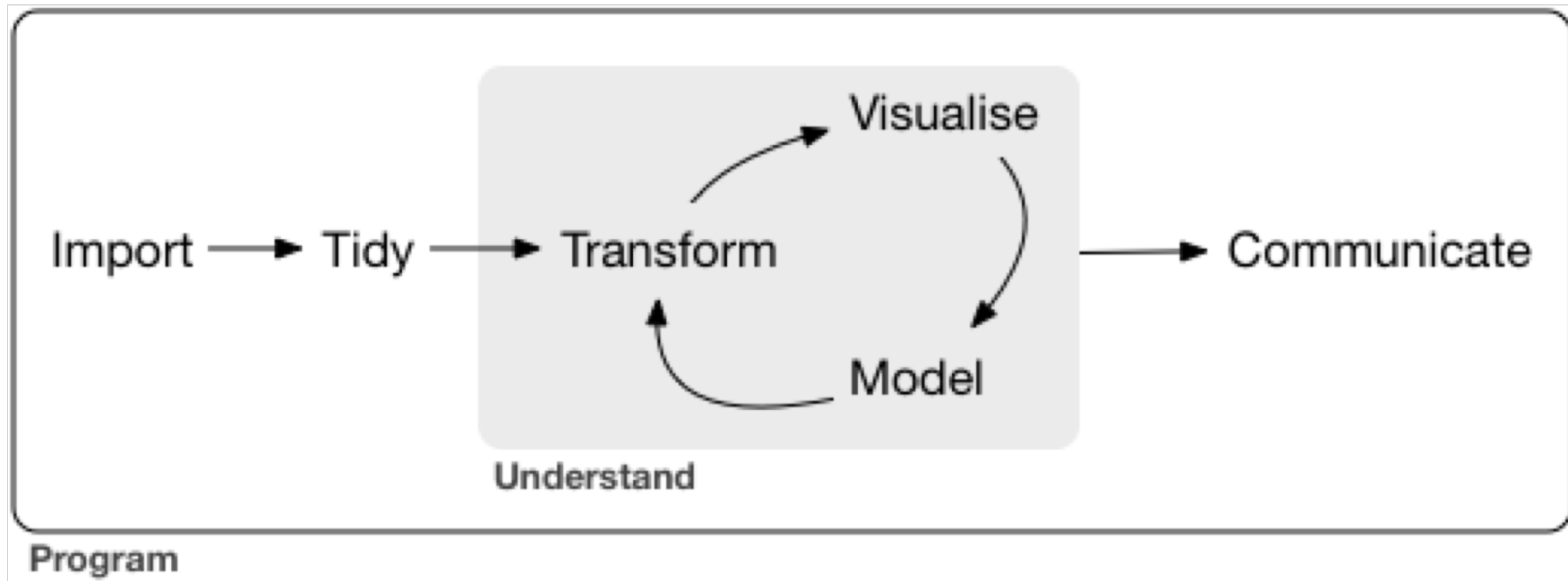
유전통계학



에너지, IoT



R 사용자 (2000년~)



- 불러오기(Import)
- 정리하기, 타이디하게 하기(tidy)
- 변형하기(transform)
- 시각화하기(visualize)
- 모델링하기(model)
- 소통하기(communicate)

Tidy 데이터

country	year	cases	population
Afghanistan	1999	745	19987071
Afghanistan	2000	2666	20595360
Brazil	1999	37737	172006362
Brazil	2000	80488	174604898
China	1999	212258	127291272
China	2000	213766	128042583

변수

country	year	cases	population
Afghanistan	1999	745	19987071
Afghanistan	2000	2666	20595360
Brazil	1999	37737	172006362
Brazil	2000	80488	174604898
China	1999	212258	127291272
China	2000	213766	128042583

관측값

country	year	cases	population
Afghanistan	99	745	19987071
Afghanistan	00	2666	20595360
Brazil	99	37737	172006362
Brazil	00	80488	174604898
China	99	212258	127291272
China	00	213766	128042583

값

Tidy 하지 않은 데이터

```
table2
#> # A tibble: 12 x 4
#>   country      year type      count
#>   <chr>      <int> <chr>    <int>
#> 1 Afghanistan 1999 cases         745
#> 2 Afghanistan 1999 population 19987071
#> 3 Afghanistan 2000 cases         2666
#> 4 Afghanistan 2000 population 20595360
#> 5 Brazil      1999 cases         37737
#> 6 Brazil      1999 population 172006362
```

```
table3
#> # A tibble: 6 x 3
#>   country      year rate
#>   * <chr>      <int> <chr>
#> 1 Afghanistan 1999 745/19987071
#> 2 Afghanistan 2000 2666/20595360
#> 3 Brazil      1999 37737/172006362
#> 4 Brazil      2000 80488/174504898
#> 5 China      1999 212258/1272915272
#> 6 China      2000 213766/1280428583
```

```
# 티블 두 개로 나누어짐
table4a # cases
#> # A tibble: 3 x 3
#>   country      `1999` `2000`
#> * <chr>      <int> <int>
#> 1 Afghanistan     745    2666
#> 2 Brazil          37737  80488
#> 3 China           212258 213766
table4b # population
#> # A tibble: 3 x 3
#>   country      `1999`      `2000`
#> * <chr>      <int>      <int>
#> 1 Afghanistan 19987071 20595360
#> 2 Brazil      172006362 174504898
#> 3 China      1272915272 1280428583
```

데이터 사이언스 도구

1. 데이터프레임 기반: python pandas, spark dataframe

Announcement: DataFrame-based API is primary API

The MLlib RDD-based API is now in maintenance mode.

As of Spark 2.0, the [RDD](#)-based APIs in the `spark.mllib` package have entered maintenance mode. The primary Machine Learning API for Spark is now the [DataFrame](#)-based API in the `spark.ml` package.

데이터 사이언스 도구

1. 데이터프레임 기반: python pandas, spark dataframe

2. pipe/pipeline

```
foo_foo <- hop(foo_foo, through = forest)
foo_foo <- scoop(foo_foo, up = field_mice)
foo_foo <- bop(foo_foo, on = head)
```

VS

```
foo_foo %>%
  hop(through = forest) %>%
  scoop(up = field_mice) %>%
  bop(on = head)
```

- 디버깅이 쉽다
- 동사(함수)가 주목된다
- 읽기가 쉽다

데이터 사이언스 도구

1. 데이터프레임 기반: python pandas, spark dataframe
2. pipe/pipeline
3. “많은 주제를 넓고 얇게 살펴보기다는 깊게 파면 더 빨리 할 수 있을 것이다.”

빅데이터와 R

- "전체 데이터는 클 수도 있지만, 특정 문제에 답을 얻는 데 필요한 데이터는 작은 경우가 많다."
- data.table, sparklyr 등

데이터 과학 컬쳐

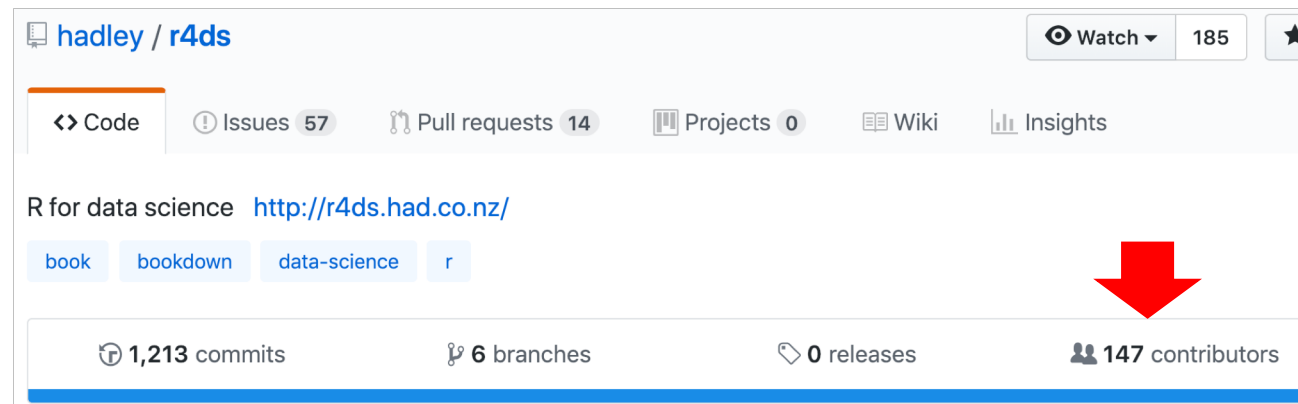
- Communication
 - "다른 사람에게 설명할 수 없다면 분석이 얼마나 훌륭한지는 중요하지 않다."
 - markdown, bookdown

데이터 과학 컬쳐

- Communication

- "다른 사람에게 설명할 수 없다면 분석이 얼마나 훌륭한지는 중요하지 않다."
- rmarkdown, bookdown

- open project for book



The screenshot shows the GitHub repository page for 'hadley / r4ds'. The repository is titled 'R for data science' with the URL <http://r4ds.had.co.nz/>. The repository has 1,213 commits, 6 branches, 0 releases, and 147 contributors. A red arrow points to the '147 contributors' count. The repository is watched by 185 people. The repository is categorized as a 'book', 'bookdown', 'data-science', and 'r' project. The repository is public and has 57 issues, 14 pull requests, and 0 projects. The repository is also linked to a Wiki and Insights page.

데이터 과학 컬쳐

- Communication

- "다른 사람에게 설명할 수 없다면 분석이 얼마나 훌륭한지는 중요하지 않다."
- rmarkdown, bookdown

- open project for book

- twitter #rstats #r4ds #Tidytuesday

The image shows a Twitter profile for 'R4DS online learning community' (@R4DScommunity). The profile has 1,649 tweets, 614 following, and 4,225 followers. The bio describes it as the Twitter home of the #R4DS Online Learning Community, inspired by the R for Data Science text, and mentions #rstats and #DataScience. Below the profile is a tweet from @noccaea (LittleSquirrel) about a #Tidytuesday challenge on space launches. The tweet includes a bar chart showing launches in different blocs (Asia, Europe, URSS) and a line chart showing launches per state bloc from 1980 to 2010.

R4DS online learning community
@R4DScommunity

Tweets 1,649 Following 614 Followers 4,225 Following

Twitter home of the #R4DS Online Learning Community (inspired by the R for Data Science text). We 🥰 #rstats & #DataScience. Join us: bit.ly/R4DSslack Slack!

LittleSquirrel @noccaea · 4h

#Tidytuesday challenge 3: Space race. The opposition between the US and the Soviet Union is flagrant, as is their decline at the end of cold war. In recent days, Asia countries have risen as the main launchers. In the west, private companies now dominate the sector. #r4ds

Space launches in different blocs?

As the market has mostly been privatised, state accounts for most launches in Russia and the states

Year of launches

Launches per state bloc
State dominated the early race, Asia is booming now

R 도움받기 - 문제 발생시

- 구글 검색
 - 오류메시지
 - 'R' 추가
- stackoverflow



R 도움받기 - 문제예방

- 블로그 (RStudio, R-bloggers)
- 트위터 (@hadleywickham, @statgarrett, @rstudiotips, #rstats, #r4ds) 팔로우

R 도움받기 - 한국어

- 페이스북 그룹
- 온라인, 오프라인 강의, 밋업
- 책
- <https://sulgik.github.io/r4ds/>