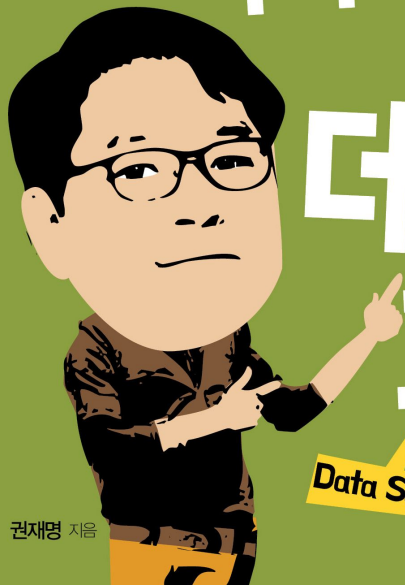


Silicon Valley

실리콘밸리
데이터과학자가
알려주는
Data Scientist

따라 하며
배우는
데이터
과학



Data Science

권재명 지음

Jpub
재이퍼블

12장. 분석 결과 정리와 공유, R마크다운

<따라 하며 배우는 데이터과학>
강의노트

2017년 8월 버전

목차

12.1	의미 있는 분석과 시각화	268
12.1.1	xkcd 지리 정보 시각화	268
12.1.2	So what(그래서 뭐)?	270
12.1.3	무쓸모 지표	270
12.1.4	의미 있는, 액서너블한 결론	271
12.2	분석의 타당성	271
12.3	보고서 작성과 구성	272
12.3.1	소통의 비결	272
12.3.2	슬라이드와 보고서의 표준적 구성	273
12.4	분석 결과의 공유	275
12.4.1	협업 도구를 활용하자	275
12.4.2	협업 도구만큼 중요한 협업 문화	276
12.4.3	코드뿐만 아니라 분석 결과도 버전 관리하자	277
12.5	R 마크다운	278
12.5.1	마크다운	278
12.5.2	분석 코드와 보고서의 결합, R 마크다운	279

분석 결과를 정리하고 공유할 때 염두에 둘 것들

1. 결과가 의미 있는가? 액서너블(actionable)한가?
2. 결과가 타당한가? 통계 방법은 정확한가? 상관 관계를 인과 관계로 오해하지는 않는가?
3. 이해하기 쉽게 쓰여졌는가? 장표와 보고서 작성을 위한 표준적인 구조를 따르는가?
4. 다른 이가 발견하기 쉽게 공유되었는가?

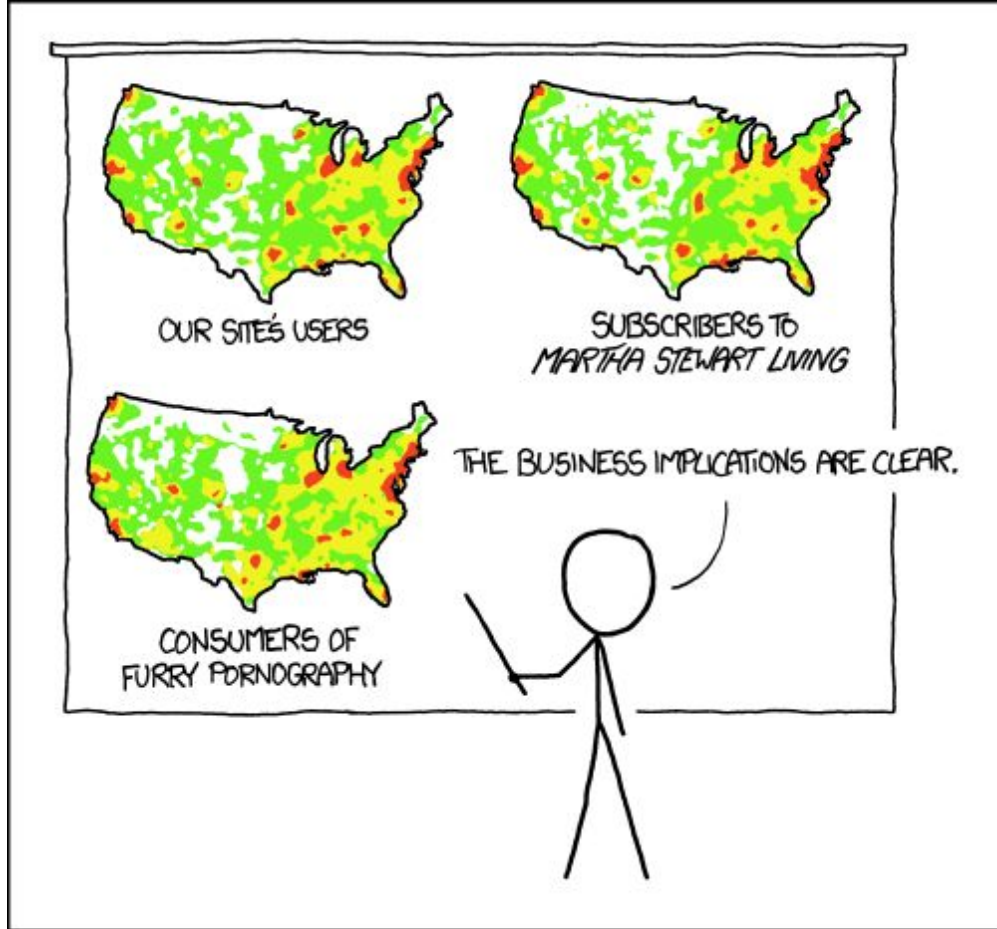
12.1 의미 있는 분석과 시각화

- 분석 결과가 의미 있는가?
- 분석 결과가 액서너블한가?
- ‘그래서 뭐?’라고 질문하면 어떻게 대답할 것인가?

<https://xkcd.com/1138/>

그림 12-1 XKCD 만화. '나를
짜증나게 하는 것 시리즈 #208.
결국 인구지도와 똑같은 지리
데이터 시각화 지도'

"보기에는 예쁘지만 의미 있는
정보를 전혀 전달하지 않는
시각화의 예"



PET PEEVE #208:
GEOGRAPHIC PROFILE MAPS WHICH ARE
BASICALLY JUST POPULATION MAPS

12.2 분석의 타당성

1. 데이터와 질문에 적절한 분석 방법이 사용되었는가?
2. 유의성을 검정하고 있다면 **P-값**을 표시하였는가?
3. 지표를 추정하고 있다면 **95%** 신뢰구간을 표시하였는가?
4. 예측 모형을 사용하고 있다면 모형의 성능을 교차검증하여 적절한 지표(**AUC**, **RMSE** 오차 등)로 요약하고 있는가?
5. 상관 관계를 인과 관계로 오해하지 않는가?

12.3 보고서 작성과 구성

1. 이해하기 쉽게 쓴다.
2. 문장은 되도록 짧고 간단하게 쓴다.
3. 사실을 확인한다.
4. 많은 사람과 나누기 전에 믿을 만한 다른 사람에게 읽어본다.

유시민 작가의 《글쓰기 특강》과 스티븐 핑커(Steven Pinker)의 《스타일(The Sense of Style)》 참조

12.4 분석 결과의 공유

12.4.1 협업 도구를 활용하자

12.4.2 협업 도구만큼 중요한 협업 문화

12.4.3 코드뿐만 아니라 분석 결과도 버전 관리하자

<https://www.youtube.com/watch?v=hNENiG7LAnc>

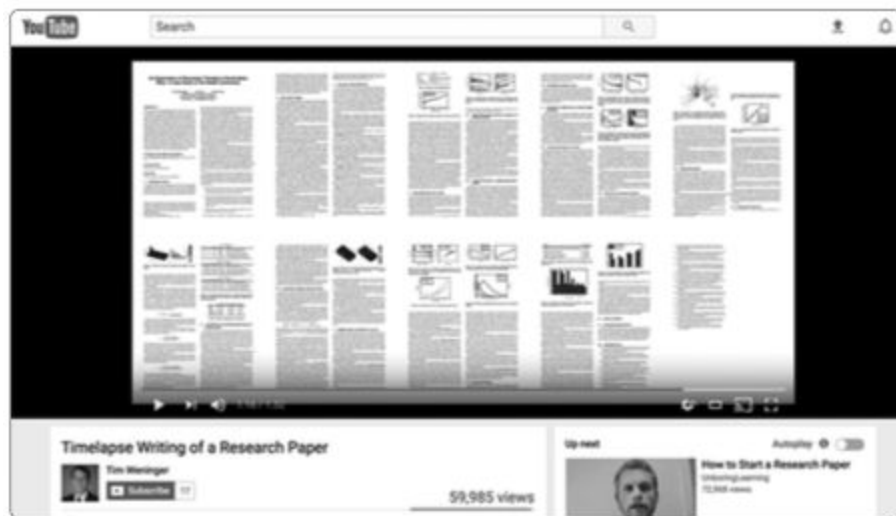


그림 12-2 연구 논문쓰기 타임랩스 비디오

출처 <https://goo.gl/L7Zz5T>

12.5 R 마크다운

12.5.1 마크다운

<http://markdownlivepreview.com/>

<https://daringfireball.net/projects/markdown/>

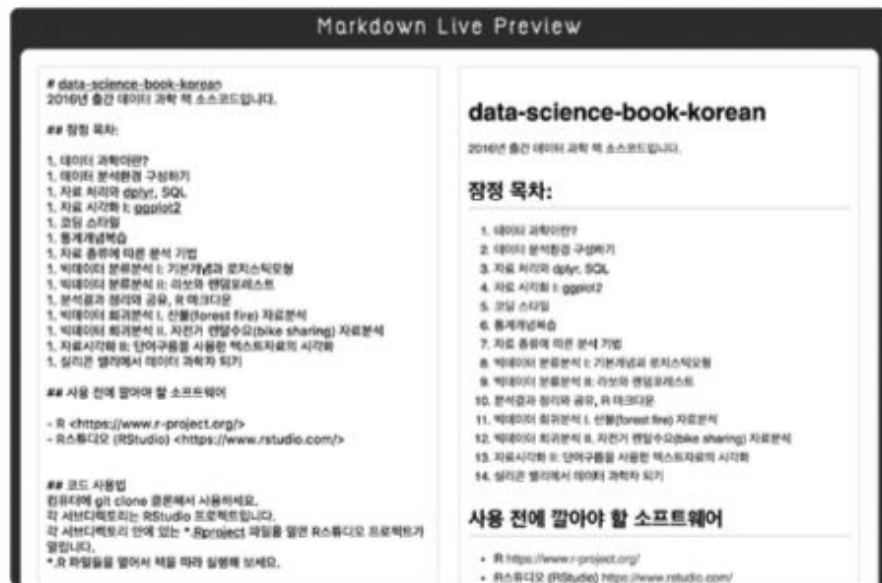


그림 12-3 마크다운 프리뷰

출처 <http://markdownlivepreview.com/>

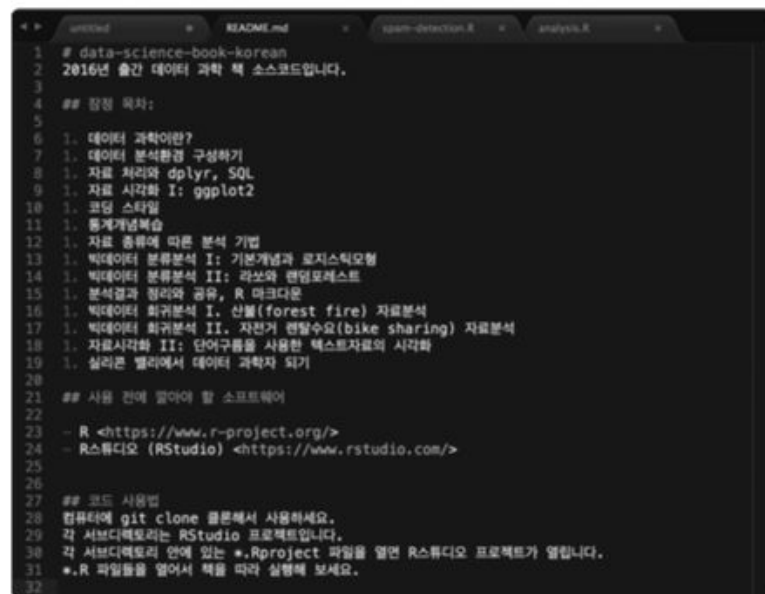


그림 12-4 서버라임에서 마크다운 포맷 지원

12.5.2 분석 코드와 보고서의 결합, R 마크다운

분석 코드 + 자료 + 보고서를 통합해서 버전관리하여

연구재현(reproducible research)를 달성.

<http://rmarkdown.rstudio.com/>

```

1 ---
2 title: "R Markdown 예제"
3 author: "권재명"
4 output: html_document
5 ---
6
7 ```{r setup, include=FALSE}
8 knitr::opts_chunk$set(echo = TRUE)
9 ```
10
11 R 마크다운을 사용하여 R 코드, 코드실행 결과 텍스트, 도표를
12 포함한 문서를 쉽게 작성할 수 있습니다.
13
14 ```{r cars, fig.width=4, fig.height=3, message=FALSE}
15 library(ggplot2)
16 qplot(speed, dist, data=cars) +
17   geom_smooth()
18 summary(cars)
19 ```

```

17:1 Chunk 2: cars R Markdown

Console

Files Plots Packages Help Viewer



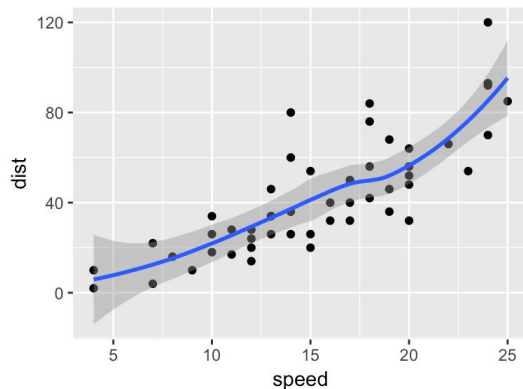
Publish

R Markdown 예제

권재명

R 마크다운을 사용하여 R 코드, 코드실행 결과 텍스트, 도표를 포함한 문서를 쉽게 작성할 수 있습니다.

```
library(ggplot2)
qplot(speed, dist, data=cars) +
  geom_smooth()
```



```
summary(cars)
```

```
##      speed      dist
##  Min.   : 4.0    Min.   : 2.00
##  1st Qu.:12.0    1st Qu.: 26.00
##  Median :15.0    Median : 36.00
##  Mean   :15.4    Mean   : 42.98
```



가장 빠르게, 가장 제대로 배우는 데이터 과학 입문서!

이 책은 '실무'에 초점을 맞춘 데이터 사이언스 '입문서'다. 다양한 배경을 가진 독자들이 가장 짧은 시간에 기본적인 데이터 사이언스 분석을 시작할 수 있도록 하였다. '가장 짧은 시간'에 배워야 하므로 필수적이지만 많은 내용은 과감히 생략하고, 설명은 최대한 간략히 하려고 노력하였다. 또한, '다양한 배경'을 가진 독자들을 위해 통계나 컴퓨터 전공 지식이 없더라도 읽을 수 있도록 하였으나, 통계의 핵심인 기초통계와 선형모형(회귀분석과 분산분석 포함)은 반드시 제대로 배울 것을 권장한다. '기본적인' 데이터 분석은 텍스트 자료, 그래프 모형, 시계열 분석, 공간자료 분석 등 개별적인 자료 형태보다는 다양한 분석에 공통적으로 적용되는 방법들을 다룬다.

이 책은 대학이나 학원의 강의 교재 혹은 자습서로 사용할 수도 있다. 강의 교재로는 학부 및 대학원 수준의 데이터 과학, 통계학, 자료분석 등의 강의에 주교재 혹은 부교재로 사용할 수 있다. 몇 주간의 단기 과정에서 일부 정만을 다루어도 좋다. R과 유닉스 코드 예를 따라 하고, 각 장 끝의 연습문제를 반드시 풀어 보도록 하자.

이 책의 대상 독자

- '데이터 사이언스 입문' 수업을 듣는 학생, 학부생 및 대학원생
- 데이터 분석 업무를 하고자 하는 관련 분야 엔지니어
- 데이터 과학 팀을 구축하고자 하는 관련 분야 매니저


요약

- 의미있는 분석과 시각화
- 분석의 타당성
- 보고서 작성과 구성
- 분석 결과의 공유
 - 협업 도구만큼 중요한 협업 문화
 - 코드 뿐 아니라 데이터, 결과도 버전 관리하자
- R 마크다운 - 코드, 데이터, 결과 보고서의 결합.

수준 

분야 컴퓨터공학 / 빅데이터

 **이것이 데이터 과학의 시작이다**
www.jpup.kr

 **아름다운재단** 출판수익 및 저작 인세의 일부는
이름다운재단의 사회공헌기금에 기부됩니다.

정가 26,000 원



9 791185 890869
ISBN 979-11-85890-86-9