



데이터 분석가의 R 패키지 개발 도전기

이영록

2021-11-19

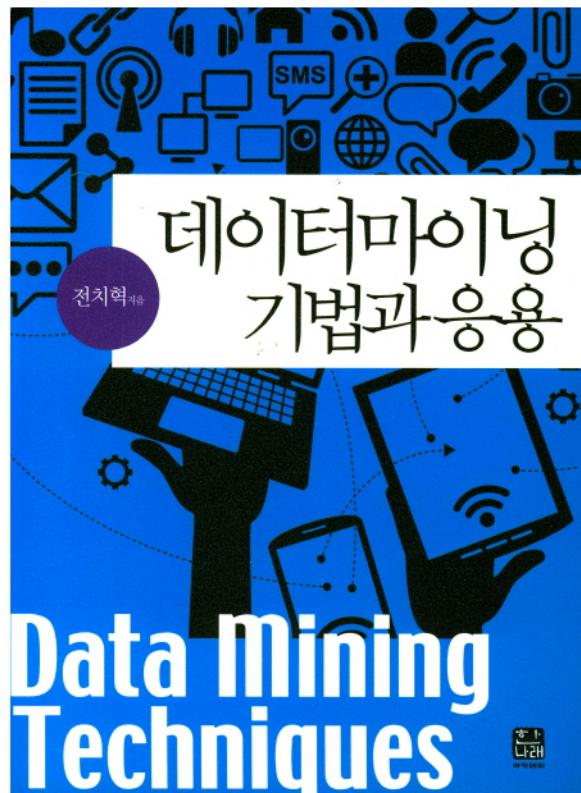
발표 개요

- 패키지 개발 배경
- 주요 작업 과정
- 데이터 분석가가 배우면 유용한 개발자 관점
- 맷음말

패키지 개발 배경

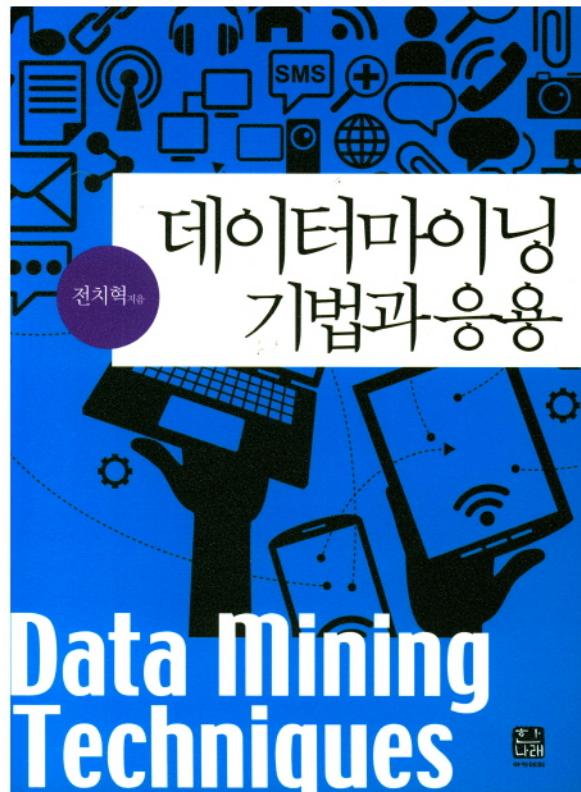
데이터마이닝 교재 참고자료 제작

데이터마이닝 교재 참고자료 제작

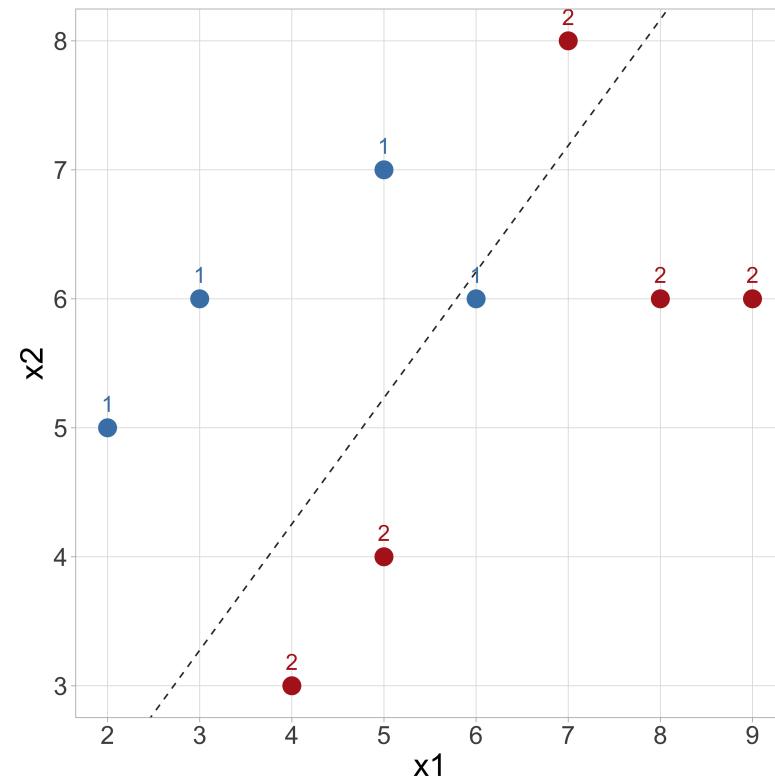


데이터마이닝 기법과 응용 / 한나래출판사 / 전치혁 저

데이터마이닝 교재 참고자료 제작

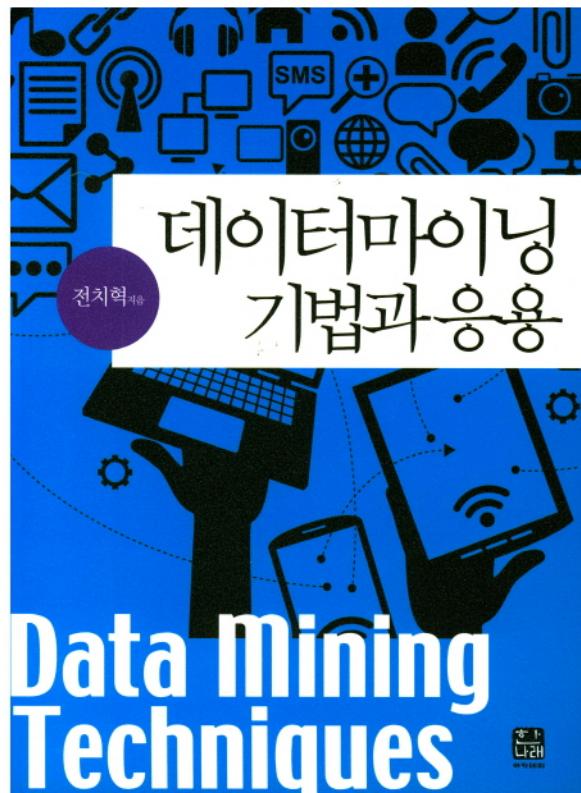


데이터마이닝 기법과 응용 / 한나래출판사 / 전지혁 저



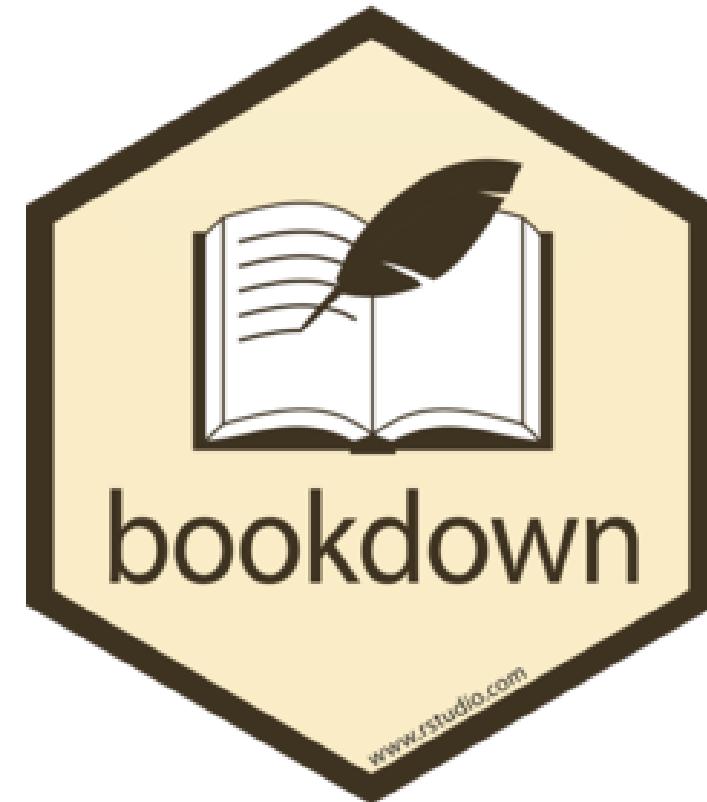
예제: 선형 분류분석

데이터마이닝 교재 참고자료 제작



데이터마이닝 기법과 응용 / 한나래출판사 / 전치혁 저

- 데이터 마이닝 이론과 알고리즘을 잘 이해할 수 있게 깔끔하게 구성된 교재
- 간단하면서 결과를 눈으로 확인할 수 있는 예제들
- 예제 결과 도출을 위한 프로그램 구현은 제공되지 않음



데이터마이닝 교재 참고자료 제작

The screenshot shows a Microsoft Word document with the title '데이터마이닝 with R' at the top. The left sidebar contains a table of contents with sections like '개요', '1 데이터마이닝 개요', '2 회귀분석', etc. The main content area has a large heading '데이터마이닝 with R' and some descriptive text. A watermark 'D' is visible in the bottom right corner.

데이터마이닝 with R

개요

1 데이터마이닝 개요

| 1부 - 예측

2 회귀분석

2.1 필요 R 패키지 설치

2.2 다중회귀모형

2.3 반응치에 대한 추정 및 예측

2.4 지시변수와 회귀모형

3 주성분분석

3.1 필요 R 패키지 설치

3.2 행렬의 분해

3.3 주성분 회귀분석

4 부분최소자승법

4.1 필요 R 패키지 설치

4.2 하나의 종속변수의 경우

4.3 다수의 종속변수의 경우

II 2부 - 분류분석

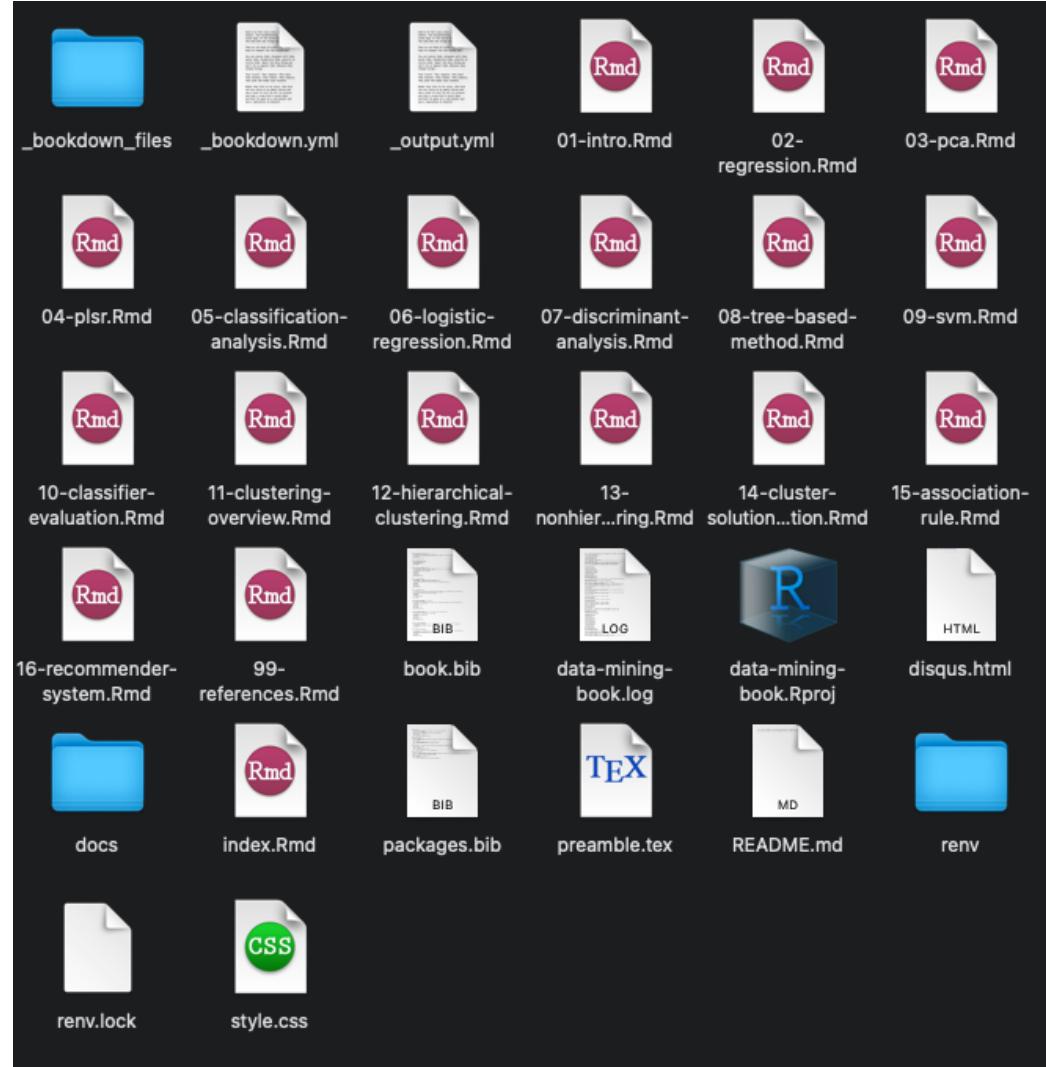
5 분류분석 개요

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js

- 온라인 버전 보조 교재로 활용
- 예제를 R로 구현한 내용을 추가
- 단순히 기존 R 패키지를 사용하는 것보다 깊이 있는 내용을 제공

데이터마이닝 with R / 전치혁, 이혜선, 이종석, 이영록 /
<http://youngroklee-ml.github.io/data-mining-book/>

데이터마이닝 교재 참고자료 제작



사용자 함수의 난립

사용자 함수의 난립



데이터 생성

함수 1

함수 2

분석

함수 3

분석

데이터 생성

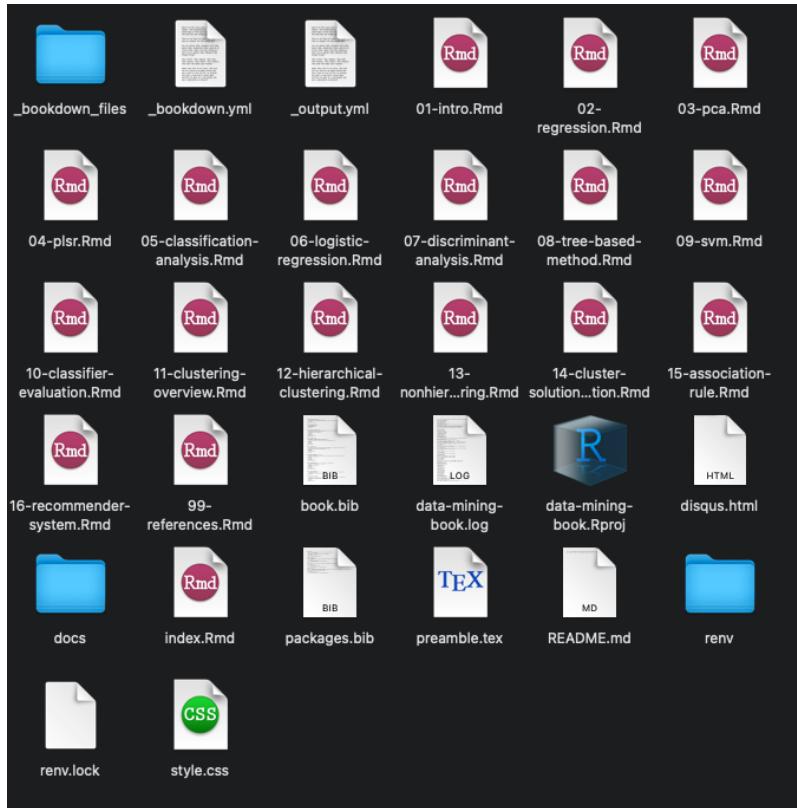
함수 4

함수 5

데이터 분석

...

사용자 함수의 난립



지금은 맞지만, 나중엔 틀릴수도?

지금은 맞지만, 나중엔 틀릴수도?

```
library(dplyr)  
  
data(binaryclass2, package = "dmtr")  
  
binaryclass2 %>%  
  select(x1, x2)
```

```
# A tibble: 9 × 2  
  x1     x2  
  <dbl> <dbl>  
1     5     7  
2     4     3  
3     7     8  
4     8     6  
5     3     6  
6     2     5  
7     6     6  
8     9     6  
9     5     4
```

```
library(dplyr)  
  
data(binaryclass2, package = "dmtr")  
  
library(MASS)  
fit ← lda(class ~ x1 + x2, binaryclass2)  
  
binaryclass2 %>%  
  select(x1, x2)
```

```
Error in select(., x1, x2): unused arguments
```

지금은 맞지만, 나중엔 틀릴수도?

{tidyverse} 0.x

```
df ← tibble(  
  a = list(c("a", "b"), "c"),  
  b = list(1:2, 3),  
  c = c(11, 22)  
)  
df %>% tidyverse::unnest(a, b)
```

```
# A tibble: 3 × 3  
  c     a     b  
  <dbl> <chr> <dbl>  
1    11 a      1  
2    11 b      2  
3    22 c      3
```

{tidyverse} 1.0.0

```
df ← tibble(  
  a = list(c("a", "b"), "c"),  
  b = list(1:2, 3),  
  c = c(11, 22)  
)  
df %>% tidyverse::unnest(c(a, b))
```

```
# A tibble: 3 × 3  
  a     b     c  
  <chr> <dbl> <dbl>  
1 a      1     11  
2 b      2     11  
3 c      3     22
```

지금은 맞지만, 나중엔 틀릴수도?

- 오류 발생 시 수정에 많은 시간 소요
- 코드 실행 오류는 없으나 결과가 틀린 경우는 오류 발견 자체가 어려움

주요 작업 과정



R Packages / O'Reilly / Hadley Wickham



R Markdown 문서 해제 및 재구성

R Markdown 문서 해체 및 재구성

```
```{r}
binaryclass2 ← tribble(
 ~id, ~x1, ~x2, ~class,
 1, 5, 7, 1,
 2, 4, 3, 2,
 3, 7, 8, 2,
 4, 8, 6, 2,
 5, 3, 6, 1,
 6, 2, 5, 1,
 7, 6, 6, 1,
 8, 9, 6, 2,
 9, 5, 4, 2
)

group_mean ← function(...) { ... }
pooled_variance ← function(...) { ... }
fisher_ld ← function(...) { ... }

w_hat ← fisher_ld(
 binaryclass2, class, x1:x2)
print(w_hat)
```
```

R Markdown 문서 해체 및 재구성

```
```{r}
binaryclass2 ← tribble(
 ~id, ~x1, ~x2, ~class,
 1, 5, 7, 1,
 2, 4, 3, 2,
 3, 7, 8, 2,
 4, 8, 6, 2,
 5, 3, 6, 1,
 6, 2, 5, 1,
 7, 6, 6, 1,
 8, 9, 6, 2,
 9, 5, 4, 2
)
```

```
group_mean ← function(...) { ... }
pooled_variance ← function(...) { ... }
fisher_ld ← function(...) { ... }

w_hat ← fisher_ld(
 binaryclass2, class, x1:x2)
print(w_hat)
```
```



R Markdown 문서 해체 및 재구성

```
```{r}
binaryclass2 ← tribble(
 ~id, ~x1, ~x2, ~class,
 1, 5, 7, 1,
 2, 4, 3, 2,
 3, 7, 8, 2,
 4, 8, 6, 2,
 5, 3, 6, 1,
 6, 2, 5, 1,
 7, 6, 6, 1,
 8, 9, 6, 2,
 9, 5, 4, 2
)

group_mean ← function(...) { ... }
pooled_variance ← function(...) { ... }
fisher_ld ← function(...) { ... }

w_hat ← fisher_ld(
 binaryclass2, class, x1:x2)
print(w_hat)
```
```



R Markdown 문서 해체 및 재구성

```
```{r}
binaryclass2 ← tribble(
 ~id, ~x1, ~x2, ~class,
 1, 5, 7, 1,
 2, 4, 3, 2,
 3, 7, 8, 2,
 4, 8, 6, 2,
 5, 3, 6, 1,
 6, 2, 5, 1,
 7, 6, 6, 1,
 8, 9, 6, 2,
 9, 5, 4, 2
)

group_mean ← function(...) { ... }
pooled_variance ← function(...) { ... }
fisher_ld ← function(...) { ... }

w_hat ← fisher_ld(
 binaryclass2, class, x1:x2)
print(w_hat)
```
```



```
```{r}
library(dmtr)

w_hat ← fisher_ld(
 binaryclass2, class, x1:x2)
print(w_hat)
```
```



함수 도움말 문서 작성



함수 도움말 문서 작성

```
#' 피셔 선형 판별 함수.  
#'  
#' 두 범주 데이터를 구분하는 피셔 선형 판별함수의  
#' 계수를 추정한다.  
#'  
#' @param .data 관측 데이터 프레임.  
#' @param .group_var 범주변수.  
#' @param .xvar 범주 분류에 사용될 변수.  
#' @return 선형 함수의 계수 벡터.  
#'  
#' @examples  
#' data(binaryclass2, package = "dmtr")  
#' fisher_ld(binaryclass2, class, c(x1, x2))  
#'  
#' @keywords discriminant-functions  
#' @export  
fisher_ld <-  
  function(.data, .group_var, .xvar) { ... }
```

함수 도움말 문서 작성

```
#' 피셔 선형 판별 함수.  
#'  
#' 두 범주 데이터를 구분하는 피셔 선형 판별함수의  
#' 계수를 추정한다.  
#'  
#' @param .data 관측 데이터 프레임.  
#' @param .group_var 범주변수.  
#' @param .xvar 범주 분류에 사용될 변수.  
#' @return 선형 함수의 계수 벡터.  
#'  
#' @examples  
#' data(binaryclass2, package = "dmtr")  
#' fisher_ld(binaryclass2, class, c(x1, x2))  
#'  
#' @keywords discriminant-functions  
#' @export  
fisher_ld <-  
  function(.data, .group_var, .xvar) { ... }
```



함수 도움말 문서 작성

```
#' 피셔 선형 판별 함수.  
#'  
#' 두 범주 데이터를 구분하는 피셔 선형 판별함수의  
#' 계수를 추정한다.  
#'  
#' @param .data 관측 데이터 프레임.  
#' @param .group_var 범주변수.  
#' @param .xvar 범주 분류에 사용될 변수.  
#' @return 선형 함수의 계수 벡터.  
#'  
#' @examples  
#' data(binaryclass2, package = "dmtr")  
#' fisher_ld(binaryclass2, class, c(x1, x2))  
#'  
#' @keywords discriminant-functions  
#' @export  
fisher_ld <-  
  function(.data, .group_var, .xvar) { ... }
```

fisher_ld {dmtr}

R Documentation

피셔 선형 판별 함수.

Description

두 범주 데이터를 구분하는 피셔 선형 판별함수의 계수를 추정한다.

Usage

```
fisher_ld(.data, .group_var, .xvar)
```

Arguments

.data 관측 데이터 프레임.

.group_var 범주변수.

.xvar 범주 분류에 사용될 변수.

Value

선형 함수의 계수 벡터.

Examples

```
data(binaryclass2, package = "dmtr")  
fisher_ld(binaryclass2, class, c(x1, x2))
```

[Package dmtr version 0.0.0.9000]

함수 도움말 작성 및 컴파일된 도움말 파일 (.Rd)

코드 테스트



코드 테스트

예제 결과 재현성 테스트

(식 7.16)에 의한 분류 경계식은 다음과 같이 주어지므로

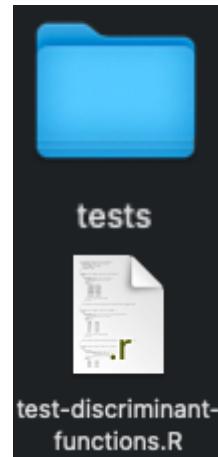
$$-1.5080X_1 + 1.5418X_2 = 0.5264 \quad (\text{또는 } X_2 = 0.3414 + 0.9781X_1)$$

코드 테스트

예제 결과 재현성 테스트

(식 7.16)에 의한 분류 경계식은 다음과 같이 주어지므로

$$-1.5080X_1 + 1.5418X_2 = 0.5264 \quad (\text{또는 } X_2 = 0.3414 + 0.9781X_1)$$



코드 테스트

예제 결과 재현성 테스트

(식 7.16)에 의한 분류 경계식은 다음과 같이 주어지므로

$$-1.5080X_1 + 1.5418X_2 = 0.5264 \quad (\text{또는 } X_2 = 0.3414 + 0.9781X_1)$$

```
library(testthat); library(dmtr);
test_that("Fisher discriminant function matches", {
  local_edition(3)
  expect_equal(
    fisher_ld(binaryclass2, class, x1:x2),
    c(x1 = -1.5080, x2 = 1.5418),
    tolerance = 1e-3,
    ignore_attr = TRUE
  )
})
```

Test passed 🎉

패키지 웹사이트 생성



패키지 웹사이트 생성



Developed by Youngrok Lee.

Site built with [pkgdown](#) 1.6.1.

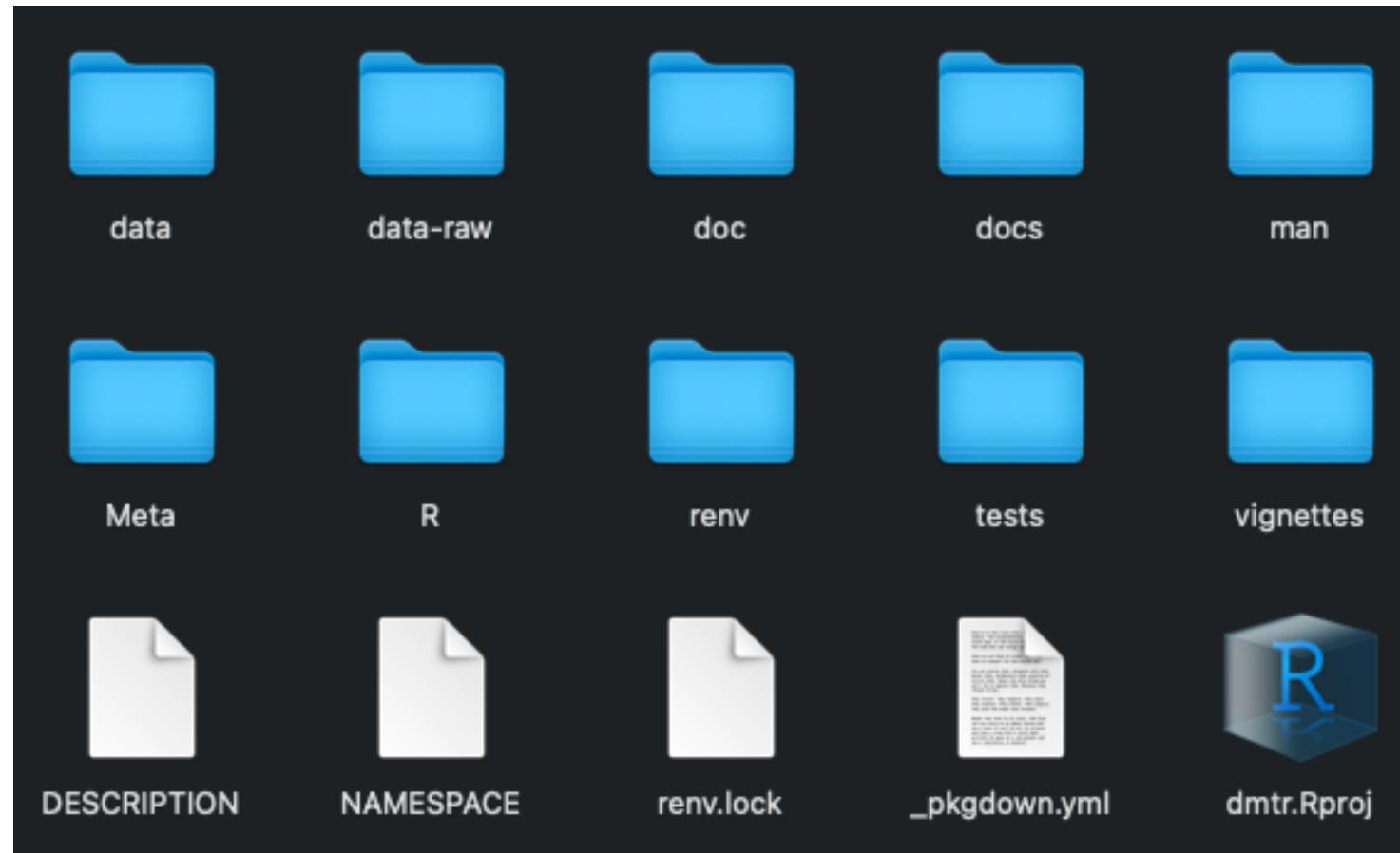
<https://youngroklee-ml.github.io/dmtr/>

데이터 분석가가 배우면 유용한 개발자 관점

체계적인 프로젝트 풀더 구조

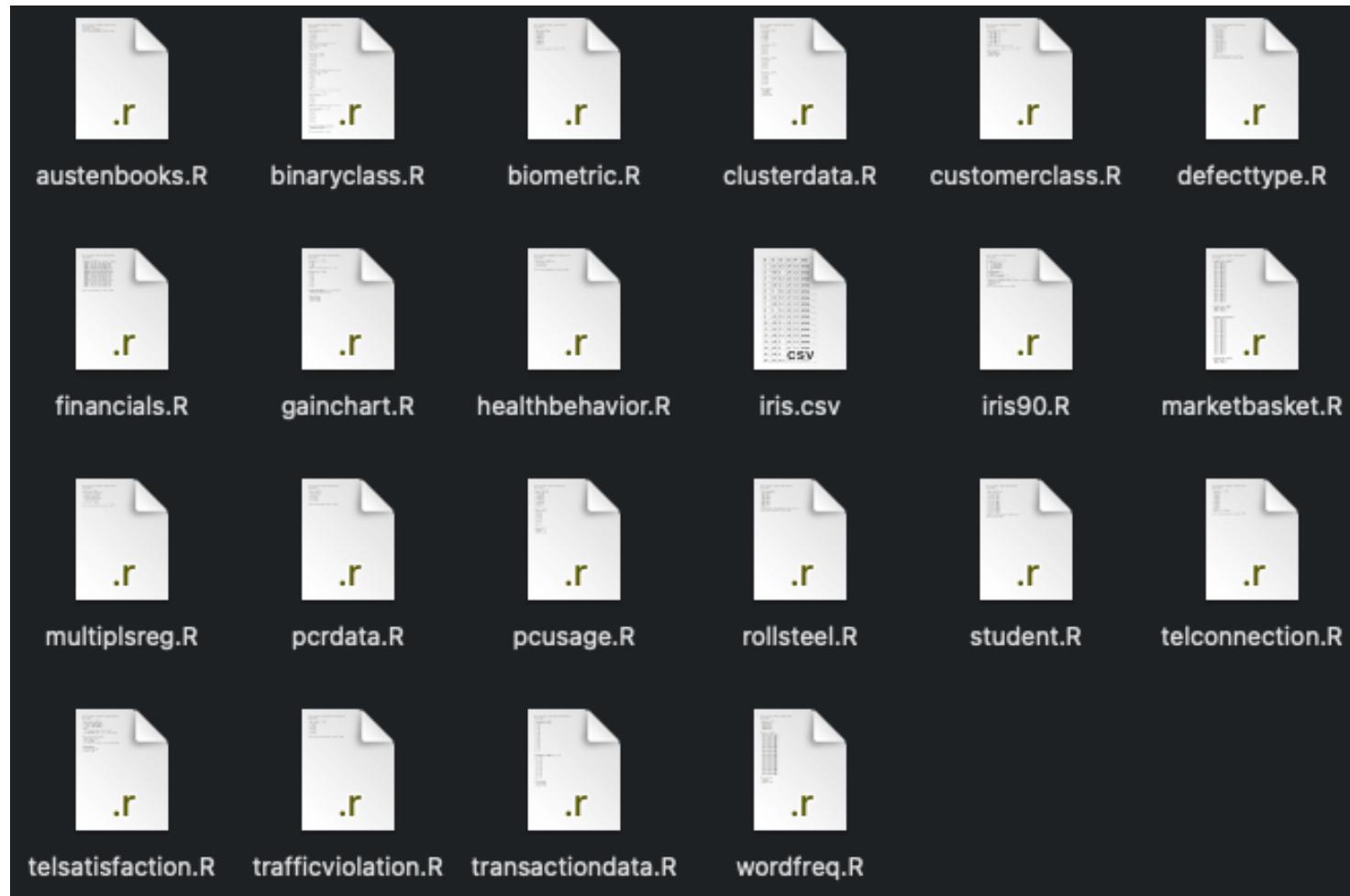
체계적인 프로젝트 폴더 구조

- 프로젝트 폴더



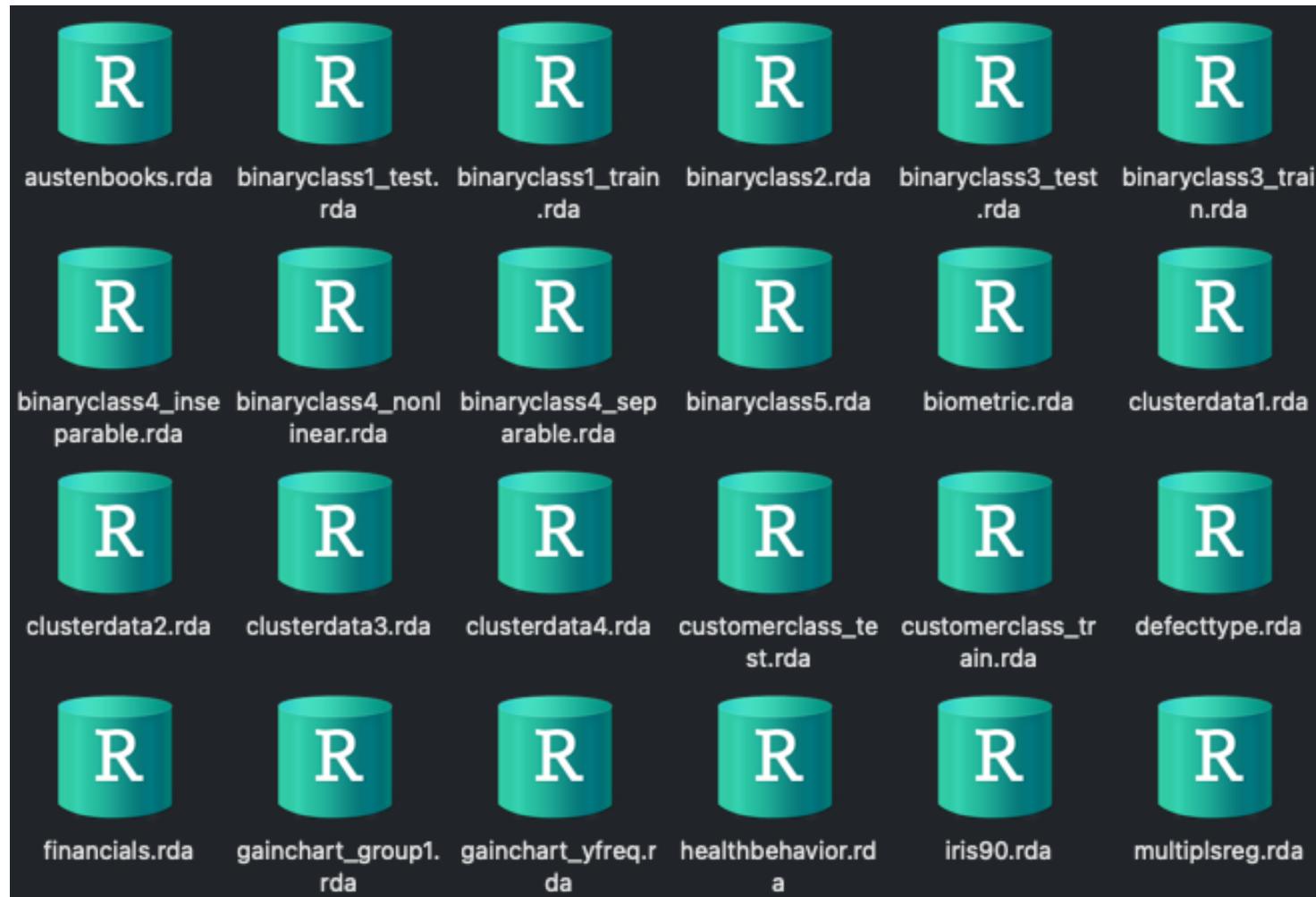
체계적인 프로젝트 폴더 구조

- data-raw/: 데이터 생성/쿼리 R 스크립트 (.R)



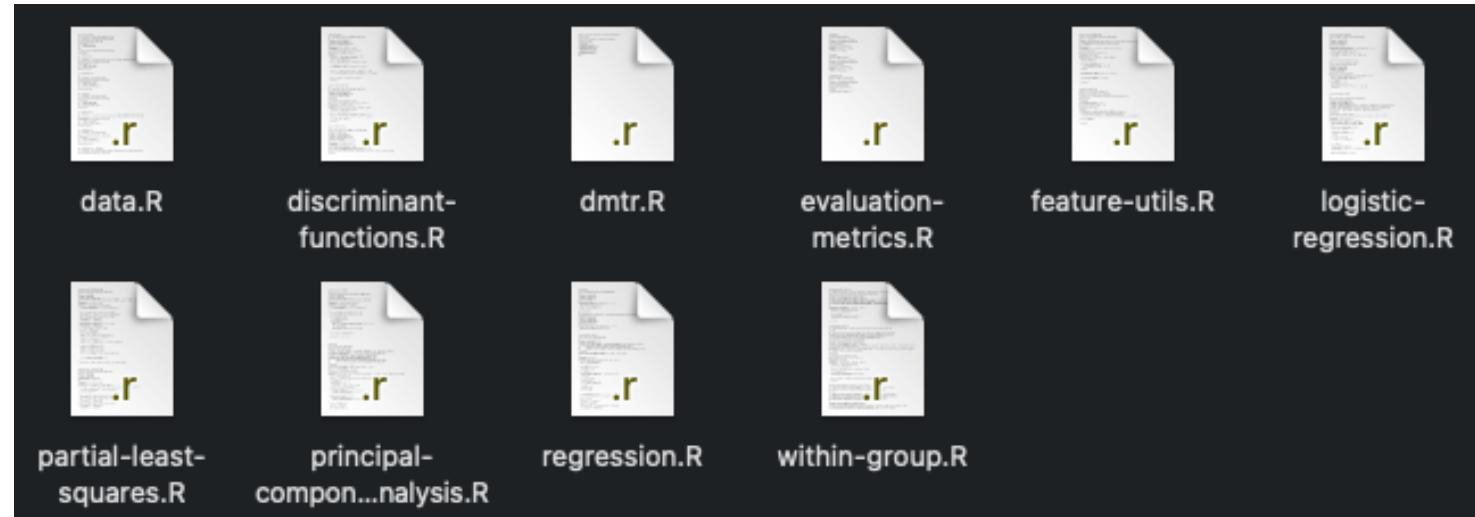
체계적인 프로젝트 폴더 구조

- data/: 생성된 데이터 파일 (.rda)



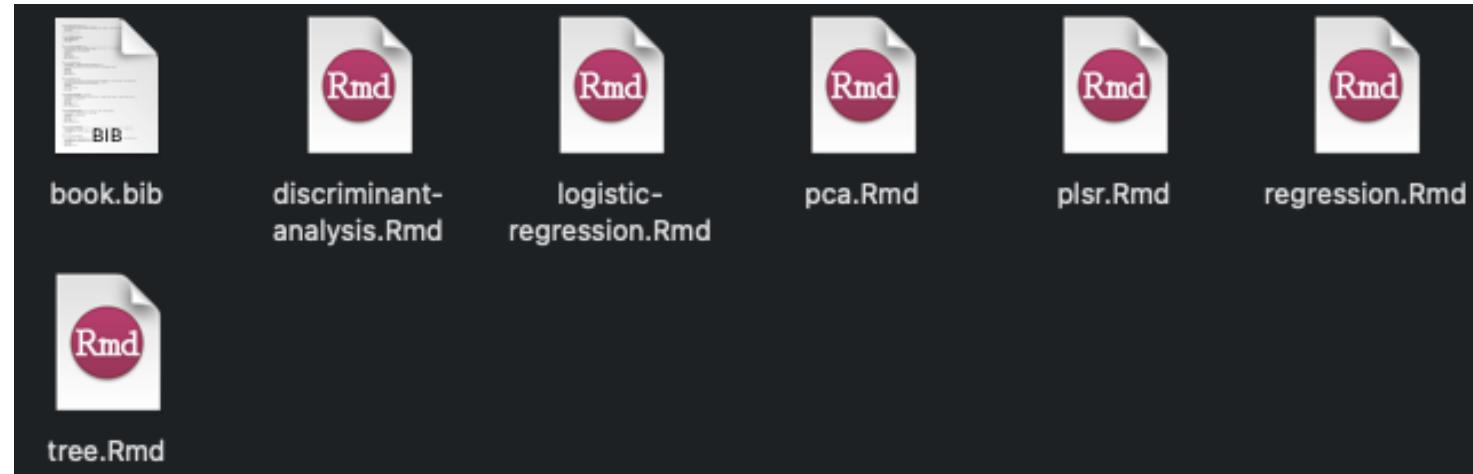
체계적인 프로젝트 폴더 구조

- R/: 재사용을 위해 모듈화된 데이터 분석 함수 (.R)



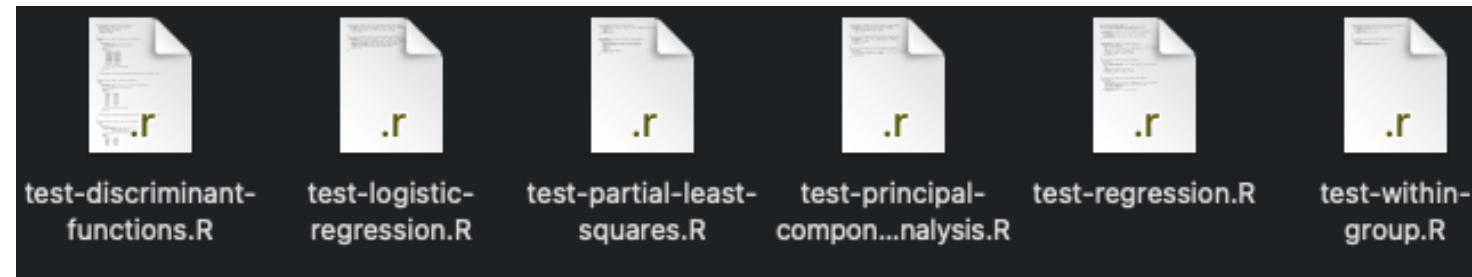
체계적인 프로젝트 폴더 구조

- vignettes/: 데이터 분석 문서 소스 (.Rmd)



체계적인 프로젝트 폴더 구조

- tests/: 코드 테스트 (.R)

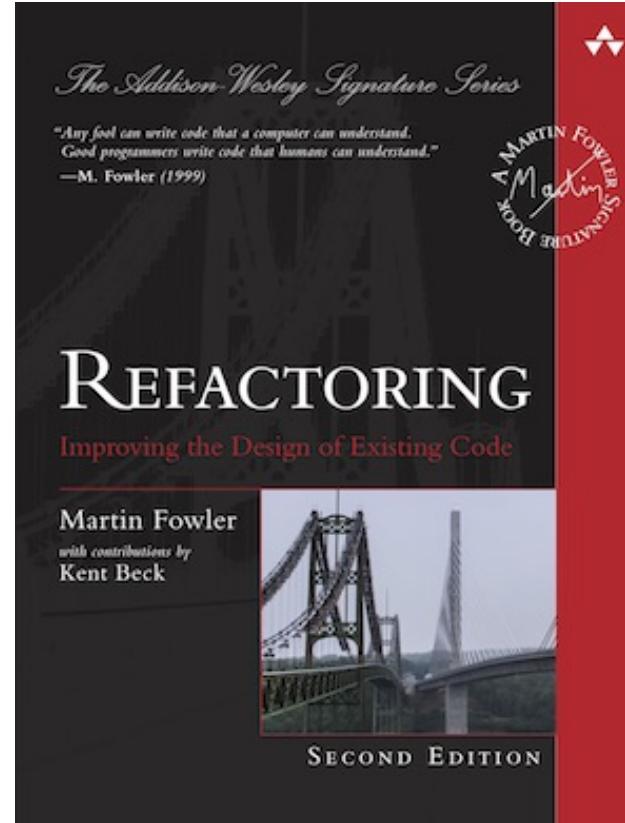


체계적인 프로젝트 폴더 구조

- 일관된 프로젝트 폴더 구조
 - **data-raw/**: 데이터 생성/쿼리 R 스크립트 (.R)
 - **data/**: 생성된 데이터 파일 (.rda)
 - **R/**: 재사용을 위해 모듈화된 데이터 분석 함수 (.R)
 - **vignettes/**: 데이터 분석 문서 소스 (.Rmd)
 - **tests/**: 코드 테스트 (.R)



리팩토링



Refactoring: improving the design of existing code / Addison-Wesley Professional / Martin Fowler

- 코드 테스트

```
==> devtools::test()

  i  Loading dmtr
  i  Testing dmtr
  ✓ |  OK F W S | Context
  ✓ |    9      | discriminant-functions [0.3 s]
  ✓ |    2      | logistic-regression
  ✓ |    2      | partial-least-squares
  ✓ |    3      | principal-component-analysis
  ✓ |    4      | regression
  ✓ |    2      | within-group

== Results ==
Duration: 0.5 s

[ FAIL 0 | WARN 0 | SKIP 0 | PASS 22 ]
```

리팩토링

- 네임 스페이스 명시화

```
library(dmtr)
library(corr)
library(dplyr)
binaryclass2 %>%
  select(x1, x2) %>%
  correlate()
```

```
# A tibble: 2 × 3
  term      x1      x2
  <chr>    <dbl>   <dbl>
1 x1        NA     0.411
2 x2     0.411 NA
```

```
library(dmtr)
library(dplyr)
binaryclass2 %>%
  select(x1, x2) %>%
  corr::correlate()
```

```
# A tibble: 2 × 3
  term      x1      x2
  <chr>    <dbl>   <dbl>
1 x1        NA     0.411
2 x2     0.411 NA
```

- 메타 프로그래밍
 - 하드코딩 지양



맺음말

데이터 분석가에게 패키지 개발이란?

데이터 분석가에게 패키지 개발이란?

- 분석 과정과 및 결과의 배포
 - 데이터, 분석 코드, 리포트 등
 - 재현가능성 검증



조리순서 *Steps*



- ➊ 양파, 당근, 애호박을 채썰어서 준비합니다.
- ➋ 비빔밥에 비벼먹을 양념장을 준비합니다.
고추장 2T +간장 2T+설탕 1.5T+깨소금, 참기름 +식초 1T를 넣고 잘섞어 양념장을 만듭니다.
- ➌ 먼저 당근을 소금 1꼬집 넣고 색깔이 나게 볶아줍니다.

데이터 분석가에게 패키지 개발이란?

- 생산성 향상을 위한 배움의 과정
 - 프로젝트 관리
 - 리팩토링



어떻게 시작할까?

어떻게 시작할까?

- 마음가짐: 위축되거나 포기하지 않기
 - 단발성 분석 업무보다 많은 시간과 노력이 소요
 - 경험의 축적이 중요
 - 자신에게 맞는 기대 수준과 페이스 찾기
- 기존에 수행한 간단한 분석 프로젝트로 시작
 - 하나의 데이터, 하나의 함수, 하나의 테스트부터 시작
 - 사용자부터 피드백 얻기
- 분석 업무와 패키지 개발의 선순환
 - 습득한 개발 방법을 분석 업무에 적용
 - 분석 업무의 유연성 확장
 - 새로운 분석 업무에서 패키지 개발의 아이디어 얻기

