

[2]데이터준비

- ✓ Data Preparation
- ✓ 데이터 수집
- ✓ Data Acquisition
- ✓ Data Processig
- ✓ 데이터 준비 R 예제

Data Preparation



- 데이터준비(Data Preparation)
- 문제에 정의된 데이터를 분석할 수 있는 형태로 만드는 과 정
 - 데이터 선택(Select Data): 분석할 문제관련 속성 선택
 - 데이터 생성(Construct Data): 기본 값으로 속성, 항목 을 추가
 생성
 - 데이터 통합(Integrate Data): 여러 소스로부터 데이터 통합
 - 데이터 형식화(Format Data): 분석방법에 적합한 형식으로 변환
 - 데이터 정제(Clean Data): 잡음(Noise), 이상치(Outlier), 누락
 값(Missing Value) 를 식별하여 처리

데이터 수집



- 문제에 정의된 관심을 가지는 현상을 데이터로 만드는 과정
 - 데이터 취득 (data acquisition, data import, data select)
 - 여러 소스로부터 데이터를 얻는 작업
 - 기존 데이터 셋 사용
 - 실험 계획에 따라 새로운 데이터 직접 수집
 - 품질 , 수집비용, 수집환경 등 고려
 - 수집할 표본(sample)개수 결정
 - 현상의 이해: 기간, 최소인원수 확보
 - 현상 일반화 : 통계적 추론의 신뢰구간으로 계산
 - 현상의 예측: 문제, 모델의 특성에 따라 표본개수 예측이 어려움, 학습할
 데이터양이 클수록 예측정확도 높아짐 (빅데이터를 이용한 예측)
 - 수집할 데이터의 속성 결정
 - 속성 측정 방법 결정-수치형, 범주형

Data Acquisition



- 데이터 취득 (data acquisition, data import, data select)
- 여러 소스로 부터 데이터를 얻는 과정
- 데이터 제공 형태
 - 표형태의 텍스트 파일(csv), 엑셀파일(xls), 관계형데이터베이스(sql), 데 이터오브젝트 형태(json, xml)
- 데이터 셋
 - R에서 제공하는 예제 데이터
 - help(package='datasets')
 - 국외
 - UCI 머신러닝 리포 [UCI Machine Learning Repository] (https://archive.ics.uci.edu/ml/index.php)
 - 머신러닝/데이터 과학 정보공유/경연 사이트 캐글 (https://www.kaggle.com/)
 - 위키피디아의 머신러닝 연구를 위한 데이터세트 리스트 (https://goo.gl/SpCOIK)
 - Gapminder World 지표 (http://www.gapminder.org/data)
 - 국내
 - 국가통계포털 (http://kosis.kr/index/index.do)
 - 공공데이터 포탈(<u>https://www.data.go.kr</u>)
 - 빅데이터센터(https://kbig.kr)

R dataset



help(package='datasets')

The R Datasets Package





Documentation for package 'datasets' version 3.4.3

DESCRIPTION file.

Help Pages

ABCDEFHIJLMNOPQRSTUVW

datasets-package The R Datasets Package

-- A --

ability.cov Ability and Intelligence Tests

<u>airmiles</u> Passenger Miles on Commercial US Airlines, 1937-1960

<u>AirPassengers</u> Monthly Airline Passenger Numbers 1949-1960

airquality New York Air Quality Measurements

anscombe Anscombe's Quartet of 'Identical' Simple Linear Regressions

<u>attenu</u> The Joyner-Boore Attenuation Data <u>attitude</u> The Chatterjee-Price Attitude Data

R dataset



"mtcars" 데이터 셋에 대한 설명

mtcars {datasets} R Documentation

Motor Trend Car Road Tests

Description

The data was extracted from the 1974 *Motor Trend* US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models).

Usage

mtcars

Format

A data frame with 32 observations on 11 variables.

- [, 1] mpg Miles/(US) gallon
- [, 2] cyl Number of cylinders
- [, 3] disp Displacement (cu.in.)
- [, 4] hp Gross horsepower
- [, 5] drat Rear axle ratio
- [, 6] wt Weight (1000 lbs)
- [, 7] qsec 1/4 mile time
- [, 8] vs V/S
- [, 9] am Transmission (0 = automatic, 1 = manual)
- [,10] gear Number of forward gears
- [,11] carb Number of carburetors

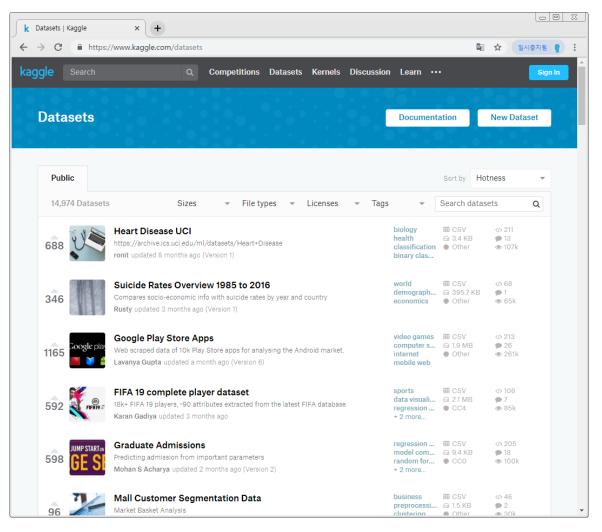
Source

Henderson and Velleman (1981), Building multiple regression models interactively. *Biometrics*, 37, 391–411.

머신러닝/데이터 과학 dataset



https://www.kaggle.com/datasets



Data Processing



- 데이터 가공(data processing)
- 각 기관으로부터 수집된 데이터를 통합하고 필요한 변수들을 테이블로 정리
- 분석을 위한 표준 테이블 형태로 변환
 - 각 행은 개별 관찰 항목(item, record, object)
 - 각 열은 개별 속성(attribute, feature, variable)
 - 각 테이블에는 단일 유형의 데이터로 구성
 - 여러 테이블이 존재하는 경우 개별 테이블을 연결할 수 있는 공통된 속성 필요
 - R에서는 JSON, CSV, XML등 널리 사용되는 형식의 파일을 테이블 형태로 불러오는 라이브러리를 제공

Name [‡]	Height [‡]	weight $^{\scriptsize \scriptsize $	sex [‡]
hong	178	78.4	male
kim	166	70.3	male
min	174	83.2	female

[개인별 키/몸무게/성별 테이블]

Data Processing



- 데이터 선택,추가,결합하기
 - 필요한 부분을 선택하고 추출하기
 - 항목 선택 및 추출
 - 속성 선택 및 추출
 - 속성 변환 및 추가
 - 속성추가
 - 자료형 변환
 - 단위변환
 - 결합하기
 - 테이블 결합

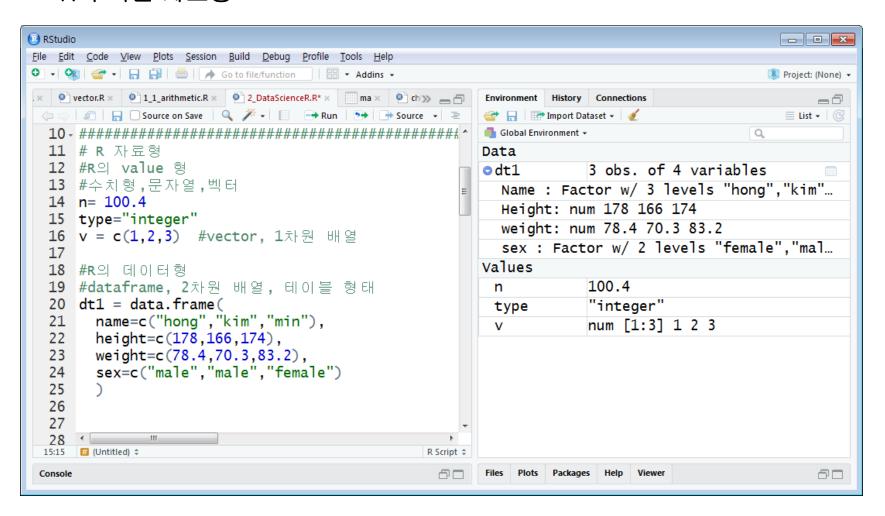
ID	password	name
1		
2		
3		

ID	age	mileage
1		
2		
3		

ID	password	name	age
1			
3			



• R의 기본 자료형





R 제공 데이터 셋 "mtcars"

:Motor Trend Car Road Tests 32개 자동차(1973 ~ 74 모델)의 설계 및 성능 특징 11개 포함 데이터

```
> str(mtcars) #데이터 셋의 특징 확인
'data.frame':
               32 obs. of 11 variables:
$ mpg : num 21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
$ cyl : num 6 6 4 6 8 6 8 4 4 6 ...
$ disp: num 160 160 108 258 360 ...
$ hp : num 110 110 93 110 175 105 245 62 95 123 ...
$ drat: num 3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
     : num 2.62 2.88 2.32 3.21 3.44 ...
$ qsec: num 16.5 17 18.6 19.4 17 ...
 $ vs : num
$ am : num
$ gear: num 4 4 4 3 3 3 3 4 4 4 ...
 $ carb: num 4 4 1 1 2
> summary(mtcars) #데이터셋의 통계 요약
                     cy1
                                     disp
Min. :10.40
                Min.
                       :4.000
                                Min. : 71.1
                                               Min. : 52.0
                1st Qu.:4.000
1st Qu.:15.43
                               1st Qu.:120.8
                                               1st Qu.: 96.5
                Median :6.000
Median :19.20
                               Median :196.3
                                               Median :123.0
Mean :20.09
                Mean :6.188
                                Mean :230.7
                                               Mean :146.7
3rd Qu.:22.80
                3rd Qu.:8.000
                                3rd Qu.:326.0
                                               3rd Qu.:180.0
       :33.90
                Max.
                       :8.000
                                       :472.0
                                               Max.
Max.
                                Max.
                                                      :335.0
     drat
                      wt
                                     gsec
       :2.760
                       :1.513
                                      :14.50
Min.
                Min.
                                Min.
1st Qu.:3.080
                1st Qu.:2.581
                                1st Qu.:16.89
                Median :3.325
Median :3.695
                               Median :17.71
Mean :3.597
                Mean
                     :3.217
                                Mean :17.85
 3rd Qu.:3.920
                3rd Qu.:3.610
                                3rd Qu.:18.90
       :4.930
                       :5.424
                Max.
                                Max.
                                       :22.90
Max.
      VS
                       am
                                       gear
Min.
       :0.0000
                 Min.
                        :0.0000
                                  Min. :3.000
1st Ou.:0.0000
                1st Ou.:0.0000
                                 1st Ou.:3.000
Median :0.0000
                 Median :0.0000
                                 Median :4.000
Mean :0.4375
                 Mean :0.4062
                                  Mean :3.688
 3rd Qu.:1.0000
                 3rd Qu.:1.0000
                                  3rd Qu.:4.000
       :1.0000
Max.
                 Max. :1.0000
                                  Max. :5.000
     carb
Min.
       :1.000
1st Qu.:2.000
Median :2.000
Mean :2.812
 3rd Qu.:4.000
Max.
       :8.000
```



```
> # 데이터 처리
> #File write / read
> write.table(mtcars, "mtcars_new.txt")
 cars = read.table("mtcars_new.txt", header=T) #헤더포함, 공백으로 속성분리
 #CSV 파일 읽기 : read.table("IrisData.csv", head=T , sep=",") #헤더포함, ","로 속성 분리
> # 데이터 살펴보기
> head(cars) # 상위 6줄 출력
                     mpg cyl disp hp drat
                                                wt qsec vs am gear
                                                                                        History
                                                                                              Connections
Mazda RX4
                           6 160 110 3.90 2.620 16.46

 List 
 □ □
                                                                               🕋 🔚 🌃 Import Dataset 🕶 💰
Mazda RX4 Wag
                   21.0
                              160 110 3.90 2.875 17.02
                                                                               Global Environment •
                                                                                                                      Q,
Datsun 710
                    22.8
                                    93 3.85 2.320 18.61
                                                                              Data
Hornet 4 Drive
                    21.4
                           6 258 110 3.08 3.215 19.44
                                                                                              32 obs. of 11 variables
                                                                              cars
Hornet Sportabout 18.7
                          8 360 175 3.15 3.440 17.02
                                                                                              3 obs. of 4 variables
                                                                              odt1
Valiant
                   18.1
                          6 225 105 2.76 3.460 20.22
                    carb
                                                                              Values
Mazda RX4
                                                                                              100.4
Mazda RX4 Wag
                                                                                             "integer"
                                                                                type
Datsun 710
                                                                                             num [1:3] 1 2 3
Hornet 4 Drive
Hornet Sportabout
Valiant
> head(cars, n=10) #상위 10줄 출력
                              disp hp drat
Mazda RX4
                           6 160.0 110 3.90 2.620 16.46
                                                                         Mazda RX4
Mazda RX4 Wag
                    21.0
                           6 160.0 110 3.90 2.875 17.02
                                                                      Mazda RX4 Wag
                                                                                                         2.875
Datsun 710
                    22.8
                           4 108.0
                                    93 3.85 2.320 18.61
                                                                         Datsun 710
                                                                                          108.0
                                                                                                93
                                                                                                         2,320
Hornet 4 Drive
                   21.4
                           6 258.0 110 3.08 3.215 19.44
                                                                       Hornet 4 Drive
                                                                                          258.0
                                                                                                    3.08
                                                                                                         3.215
Hornet Sportabout 18.7
                           8 360.0 175 3.15 3.440 17.02
                                                                     Hornet Sportabout
                                                                                                175
                                                                                                    3.15
Valiant
                   18.1
                           6 225.0 105 2.76 3.460 20.22
                                                                                                         3.460
Duster 360
                   14.3
                           8 360.0 245 3.21 3.570 15.84
                                                                         Duster 360
                                                                                 14.3
                                                                                          360.0
                                                                                                245
                                                                                                    3.21
                                                                                                         3,570
                                                                                                              15.84
Merc 240D
                   24.4
                           4 146.7 62 3.69 3.190 20.00
                                                                                          146.7
                                                                         Merc 240D
                                                                                                62
                                                                                                         3.190
                                                                                                              20.00
Merc 230
                   22.8
                                    95 3.92 3.150 22.90
                                                                          Merc 230
Merc 280
                   19.2
                           6 167.6 123 3.92 3.440 18.30
                                                                                          167.6
                                                                                                         3,440
                                                                                          167.6
                                                                                                         3.440
                                                                         Merc 280C
                                                                                                123
                                                                                                    3.92
                                                                                                              18.90
                   gear carb
                                                                         Merc 450SE
                                                                                                         4.070
                                                                                                              17.40
                                                                                                180
Mazda RX4
Mazda RX4 Wag
                                                                        Merc 450SLC
                                                                                                              18.00
Datsun 710
Hornet 4 Drive
```

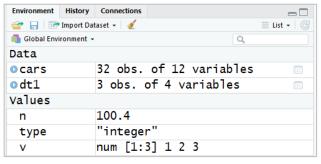
Hornet Sportabout



```
> tail(cars) # 하위 6
                mpg cyl disp hp drat
                                          wt gsec vs am gear
              26.0
                    4 120.3 91 4.43 2.140 16.7
Porsche 914-2
                     4 95.1 113 3.77 1.513 16.9
               30.4
Lotus Europa
                   8 351.0 264 4.22 3.170 14.5
Ford Pantera L 15.8
                     6 145.0 175 3.62 2.770 15.5
Ferrari Dino
              19.7
                     8 301.0 335 3.54 3.570 14.6
Maserati Bora 15.0
                     4 121.0 109 4.11 2.780 18.6
Volvo 142E
               21.4
               carb
Porsche 914-2
Lotus Europa
Ford Pantera L
Ferrari Dino
Maserati Bora
Volvo 142E
> rownames(cars) #행의 이름
[1] "Mazda RX4"
                           "Mazda RX4 Wag"
[3] "Datsun 710"
                           "Hornet 4 Drive"
 [5] "Hornet Sportabout"
                           "Valiant"
[7] "Duster 360"
                           "Merc 240D"
[9] "Merc 230"
                           "Merc 280"
[11] "Merc 280C"
                           "Merc 450SE"
[13] "Merc 450SL"
                           "Merc 450SLC"
[15] "Cadillac Fleetwood"
                           "Lincoln Continental"
[17] "Chrysler Imperial"
                           "Fiat 128"
[19] "Honda Civic"
                           "Toyota Corolla"
[21] "Toyota Corona"
                           "Dodge Challenger"
[23] "AMC Javelin"
                           "Camaro Z28"
[25] "Pontiac Firebird"
                           "Fiat X1-9"
[27] "Porsche 914-2"
                           "Lotus Europa"
[29] "Ford Pantera L"
                           "Ferrari Dino"
[31] "Maserati Bora"
                           "Volvo 142E"
> colnames(cars) #열의 이름
[1] "mpg" "cyl" "disp" "hp"
                                             "gsec" "vs"
                                "drat" "wt"
            "gear" "carb"
[9] "am"
> cars$mpg # cars 데이터의 mpg(연비) 속성 확인
[1] 21.0 21.0 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 17.8 16.4
[13] 17.3 15.2 10.4 10.4 14.7 32.4 30.4 33.9 21.5 15.5 15.2 13.3
[25] 19.2 27.3 26.0 30.4 15.8 19.7 15.0 21.4
```



```
> #모델명 속성을 추가하기
> cars$model = rownames(cars)
> rownames(cars) = NULL
> head(cars)
  mpg cyl disp hp drat wt qsec vs am gear carb model
1 21.0
           160 110 3.90 2.620 16.46 0 1
2 21.0
           160 110 3.90 2.875 17.02
3 22.8
           108 93 3.85 2.320 18.61
4 21.4 6 258 110 3.08 3.215 19.44 1 0
5 18.7
      8 360 175 3.15 3.440 17.02
           225 105 2.76 3.460 20.22
6 18.1
```



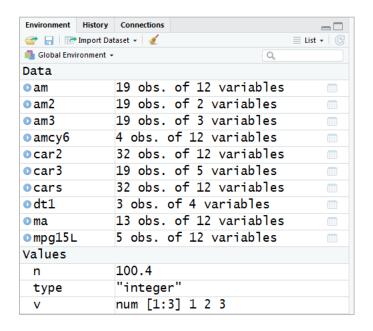
2_Dat	taScienceR	.R* ×	cars ×									
$\Leftrightarrow \Rightarrow$	2 T	Filter										Q
*	mpg [‡]	cyl [‡]	disp ‡	hp [‡]	drat [‡]	wt ÷	qsec ‡	vs [‡]	am [‡]	gear ‡	carb ‡	model [‡]
1	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4	Mazda RX4
2	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4	Mazda RX4 Wag
3	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1	Datsun 710
4	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1	Hornet 4 Drive
5	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2	Hornet Sportabout
6	18.1	6	225.0	105	2.76	3.460	20.22	1	0	3	1	Valiant
7	14.3	8	360.0	245	3.21	3.570	15.84	0	0	3	4	Duster 360
8	24.4	4	146.7	62	3.69	3.190	20.00	1	0	4	2	Merc 240D
9	22.8	4	140.8	95	3.92	3.150	22.90	1	0	4	2	Merc 230
10	19.2	6	167.6	123	3.92	3.440	18.30	1	0	4	4	Merc 280
11	17.8	6	167.6	123	3.92	3.440	18.90	1	0	4	4	Merc 280C
12	16.4	8	275.8	180	3.07	4.070	17.40	0	0	3	3	Merc 450SE
13	17.3	8	275.8	180	3.07	3.730	17.60	0	0	3	3	Merc 450SL
14	15.2	8	275.8	180	3.07	3.780	18.00	0	0	3	3	Merc 450SLC

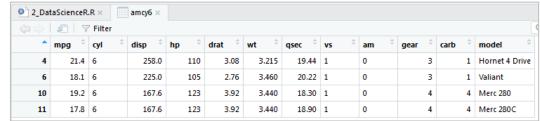


```
#행추출
mpg15L = subset(cars,mpg < 15) #연비 15미만인 데이터 추출
am = subset(cars,am == 0) #자동 변속 (automatic)만 추출
ma = cars[cars$am == 1,] #조건 추출, 수동변속만 추출
amcy6 = am[am$cyl == 6,] #조건 추출, 실린더 6개인 데이터만 추출

#열추출
am2 = am[,c(8,10)] #8,10 열추출
am3 = am[,c(1,2,3)] #1,2,3 열추출

#데이터 결합
car2 = rbind(am,ma) #행단위 결합
car3 = cbind(am3,am2) #열단위 결합
```





2_Da	2_DataScienceR.R ×		amcy6 ×	ma⇒	car3 ×	
⟨□ □⟩ Ø□ ▼ Filter						
•	mpg [‡]	cyl [‡]	disp ‡	vs ÷	gear [‡]	
4	21.4	6	258.0	1	3	
5	18.7	8	360.0	0	3	
6	18.1	6	225.0	1	3	
7	14.3	8	360.0	0	3	
8	24.4	4	146.7	1	4	
9	22.8	4	140.8	1	4	