주요 데이터셋

R 내장 데이터 셋

- data()
- data(package="datasets") : datasets 패키지에 있는 데이터셋 목록 조회
- ?데이터셋명 : 데이터셋에 대한 상세 설명 표시

패키지	데이터셋	상세
datasets	ChickWeight	
datasets	airquality	
datasets	iris	
ggplot2	diamonds	
kohonen	vintages	
kohonen	wines	
arules	AuditUCI	
googleVis	Export	GEO
plyr	baseball	시계열
TTR	ttrc	시계열
datasets	AirPassengers	시계열

패키지	데이터셋	상세
arules	Epub	트랜잭션
arules	Audit	트랜잭션
tm	acq	텍스트 문서
tm	crude	텍스트 문서
tm	txt	텍스트 문서
NetData	kracknets	Social Network
Grey's Anatomy Network of Sexual Relations http://www.babelgraph.org/data/ga_edgelist.cs v		
sentimentData, sentimentDictionary : 감성 분석		

ChickWeight datasets

- 다이어트 방법에 따른 닭의 몸무게 변화

Sefl	항목명	종류	상세
1	weight	num	▪ 몸무게
2	Time	num	▪ 시간
3	Chick	Ord.factor	▪ 1 ~ 50 ▪ 닭의 고유 ID
4	Diet	Factor	▪ 다이어트 방법 ▪ 1, 2, 3, 4

airquality in datasets

- 뉴욕의 대기 측정값

Sefl	항목명	종류	상세
1	Ozone	int	▪ 오존 농도, ppb
2	Solar.R	int	▪ 태양, lang
3	Wind	num	▪ 바람, mph
4	Temp	int	▪ 기온, F
5	Month	int	▪ 월 ▪ 1 ~ 12
6	Day	int	▪ 일 ▪ 1 ~ 31

데이터셋에서 NA를 포함한 레코드 삭제

```
data <- airfluality[complete.cases(airfluality), ]
```

iris in datasets

- 붓꽃의 3가지 종에 대해 꽃받침, 꽃잎의 길이를 정리한 데이터 150개

Sefl	항목명	종류	상세
1	Sepal.Length	num	▪ 꽃받침 길이
2	Sepal.Width	num	▪ 꽃받침 넓이
3	Petal.Length	num	▪ 꽃잎 길이
4	Petal.Width	num	▪ 꽃잎 넓이
5	Species	Factor	▪ 종 ▪ setosa, versicolor, virginica

diamonds in ggplot2

- 다이아몬드의 크기와 무게, 컷팅, 색상 등에 따른 가격

Sefl	항목명	종류	상세
1	carat	num	▪ 무게
	cut	Ord.factor	▪ 컷팅 ▪ Fair, Good, Very Good, Premium, ideal
2	color	Ord.factor	▪ 색상 ▪ J. 가장 나쁨, D. 가장 좋음
3	clarity	Ord.factor	▪ 깨끗함 ▪ I1, SI1, SI2, VS1, VS2, VVS1, VVS2, IF
4	depth	num	▪ 깊이 비율, z / mean(x, y)
5	table	num	▪ 폐놀
7	price	int	▪ 플라보노이드
8	x	num	▪ 길이, mm
9	y	num	▪ 너비, mm
10	z	num	▪ 깊이, mm

vintages in kohonen

- Wine이 만들어진 연도 또는 특정 연도나 지역에서 생산된 포도주의 종류

Sefl	항목명	종류	상세
1	vintages	Factor	▪ Wine의 Vintages ▪ Barolo, Grignolino, Barbera

wines in kohonen

- vintages에 매핑되는 포도주의 속성

Sefl	항목명	종류	상세
1	alcohol	num	▪ 알코올 도수
2	malic acid	num	▪ 사과산
3	ash	num	▪ 재
4	ash alkalinity	num	▪ 알카리도
5	magnesium	num	▪ 마그네슘
6	tot. phenols	num	▪ 페놀
7	flavonoids	num	▪ 플라보노이드
8	non-flav. phenols	num	▪ 비 flav 페놀
9	proanth	num	▪ 프로안토시아니딘
10	col. int.	num	▪ 색 선명도
11	col. hue	num	▪ 색조
12	OD ratio	num	▪ OD 비율
13	proline	num	▪ 프롤린

Exports in gooGleVis

- 국가별 수출에 따른 수익과 온라인 유무

Sefl	항목명	종류	상세
1	Country	Factor	▪ 국가
2	Profit	num	▪ 수익
3	Online	logic	▪ 온라인 유무 ▪ TRUE, FALSE

baseball in plyr(1/2)

- 메이저 리그 선수들의 년간 배팅 정보 (http://www.baseball-databank.org/)

Sefl	항목명	종류	상세
1	id	chr	▪ 선수 아이디
2	year	int	▪ 년도
3	stint	int	
4	team	chr	▪ 팀
5	lg	chr	▪ league
6	g	int	▪ 게임 수
7	ab	int	▪ 배팅 횟수
8	r	int	▪ 주루 횟수
9	h	int	▪ hits 수
10	x2b	int	▪ 2루타 수
11	x3b	int	▪ 3루타 수
12	hr	int	▪ 홈런 수
13	rbi	int	▪ 번트 수
14	sb	int	▪ 도루 수
15	cs	int	▪ caught stealing

baseball in plyr(2/2)

Sefl	항목명	종류	상세
16	bb	int	▪ 4볼 수
17	so	int	▪ 스트라이크 아웃 수
18	ibb	int	▪ intentional base on balls
19	hbp	int	▪ hits by pitch
20	sh	int	▪ sacrifice hits
21	sf	int	▪ sacrifice flies
22	gidp	int	▪ ground into double play

ttrc in TTR

- 1985년 1월 2일부터 2006년 12월 31일까지 랜덤하게 생성된 시계열 데이터

Sefl	항목명	종류	상세
1	Date	Date	▪ Format : yyyy-mm-dd
2	Open	num	
3	High	num	
4	Low	num	
5	Close	num	
6	Volumn	num	

AirPassengers in datasets

- 1949년부터 1960년까지 "월별 항공기 승객 수"에 대한 시계열 데이터

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1949	112	118	132	129	121	135	148	148	136	119	104	118
1950	115	126	141	135	125	149	170	170	158	133	114	140
1951	145	150	178	163	172	178	199	199	184	162	146	166
1952	171	180	193	181	183	218	230	242	209	191	172	194
1953	196	196	236	235	229	243	264	272	237	211	180	201
1954	204	188	235	227	234	264	302	293	259	229	203	229
1955	242	233	267	269	270	315	364	347	312	274	237	278
1956	284	277	317	313	318	374	413	405	355	306	271	306
1957	315	301	356	348	355	422	465	467	404	347	305	336
1958	340	318	362	348	363	435	491	505	404	359	310	337
1959	360	342	406	396	420	472	548	559	463	407	362	405
1960	417	391	419	461	472	535	622	606	508	461	390	432

AdultUCI in arules

- 48,842건의 인구 통계 데이터

Sefl	항목명	종류	상세
1	age	int	▪ 나이
2	workclass	Factor	▪ 직업 분류
3	fnlwgt	int	▪ 샘플링 웨이트
4	education	Ord.factor	▪ 교육 수준
5	education-num	int	▪ 교육 번호
6	marital-status	Factor	▪ 결혼 여부 및 상태
7	occupation	Factor	▪ 직업
8	relationship	Factor	▪ 관계
9	rece	Factor	▪ 인종
10	sex	Factor	▪ 성별 (Female, Male)
11	capital-gain	int	▪ 자본 이득
12	capital-loss	int	▪ 자본 손실
13	hours-per-week	int	▪ 주당 근무시간
14	native-country	Factor	▪ 태어난 나라
15	income	Ord.factor	▪ 수입 (small, large)

Epub in arules & Adult in arules

Epub in arules

2003년부터 2008년까지 비엔나대학에서 다운로드된 전자책 관련 정보

항목	상세
transactionID	▪ 사용자가 접속한 세션 아이디 ▪ 15,729 Transactions
itemsetID	
items	▪ 세션에서 다운로드한 도서 목록
TimeStamp	▪ 다운로드한 시간 (2003-03-22 12:17:31)

Adult in arules

48,842건의 인구 통계 데이터

항목	상세
transactionID	▪ 1부터 증가하는 숫자
itemsetID	
items	▪ 세션에서 다운로드한 도서 목록
TimeStamp	▪ 다운로드한 시간 (2003-03-22 12:17:31)

패키지의 텍스트 문서

- R의 Package 설치 폴더
  - C:/Program Files/R/R-x.x.x/library/
  - C:/Users/사용자_아이디/Documents/R/win-library/x.x/
- acfl in tm
  - 50개의 로이터 기사 (XML 형태)
  - R의_Package_설치_폴더/tm/texts/acfl/ 폴더에 있는 파일
- crude in tm
  - 20개의 로이터 기사 (XML 형태)
  - R의_Package_설치_폴더/tm/texts/crude/ 폴더에 있는 파일
- txt in tm
  - VCorpus 형태의 데이터셋이 아니라 파일로만 존재합니다.
  - R의_Package_설치_폴더/tm/texts/txt/ 폴더에 있는 5개의 text 파일