R 수업관련 Chapter별 데이터 셋 설명

chap06_DataVisualization

데이터 셋	Severity_Counts 데이터 셋에 관한 설명
Severity_Counts는 RSADBE 패키지에서 제공하는 데이터 셋으로 다음과 같이 소프트웨어 발표 전과 후의 버그를 측정한 10개의 벡터 자료를 제공한다. Bugs.BR/AR NT.BR/AR Major.BR/AR Critical.BR/AR H.BR/AR 단순버그 사소하지않음 중대한버그 결정적인 버그 시급한 버그	

데이터 셋	Bug_Metrics_Software 데이터 셋에 관한 설명
Bug_Metrics_Software는 RSADBE 패키지에서 제공되는 데이터 셋으로 5개의 소프트웨어 별로 발표 전과 후 버그 측정 결과를 3차원 배열구조로 데이터를 제공한다. 1면에는 소프트웨어 발표 전(Before) 버그 측정 결과, 2면에는 소프트웨어 발표 후(After) 버그 측정 결과를 제공한다.	

데이터 셋	galton 데이터 셋에 관한 설명
galton은 psych 패키지에서 제공되는 데이터 셋으로 갈턴(Galton)에 의해서 연구된 부모와 자식의 키 사이의 관계를 기록한 데이터를 제공한다. 전체 관측치는 928개 이며, 2개의 변수(child와 parent)를 제공한다. 프랜시스 칼턴(Francis Galton)은 영국 유전학자로 우생학 창시자, 종의 기원를 저술한 찰스 다윈(Darwin)의 사촌이다. 우생학이란 유전학, 의학, 통계학을 기초로 우수 유전자 증대를 목적으로 한 학문이다.	

데이터 셋	iris 데이터 셋에 관한 설명
R에서 제공되는 기본 데이터 셋으로 3가지 꽃의 종류별로 50개씩 전체 150개의 관측치로 구성된다. iris는 붓꽃에 관한 데이터를 5개의 변수로 제공하며, 각 변수의 내용은 다음과 같다. Sepal.Length(꽃받침 길이), Sepal.Width(꽃받침 너비), Petal.Length(꽃잎 길이), Petal.Width(꽃잎 너비), Species(꽃의 종류) : 3가지 종류별 50개(전체 150개 관측치)	

chap07_DataPreprocessing

dataset.csv 데이터 셋 변수 구성

변수	resident	gender	job	age	position	price	survey
척도[1]	명목	명목	명목	비율	서열	비율	등간
범위	1~5	1, 2	1~3	20~69	1~5	2.1~7.9	1~5
설명	거주시	성별	직업	나이	직위	구매금액	만족도

chap08_DataReshape

데이터 셋	hflights 데이터 셋에 관한 설명
2011년도 미국 휴스턴에서 출발하는 모든 비행기의 이륙과 착륙 정보가 기록된 것으로 227,496건의 관측치와 21개의 칼럼으로 구성된 데이터 셋이다. 주요 변수 : Year(년), Month(월), DayofMonth(일), DayOfWeek(요일), AirTime(비행시간), DepTime(출발시간), ArrTime(도착시각) TailNum(항공기 일련번호), DepDelay(출발지연시간), ArrDelay(도착지연시각), Distance(비행거리)	

데이터 셋	Indometh 데이터 셋에 관한 설명
항염증제(Indomethacin)에 대한 약물동태학에 관한 데이터 셋으로 약물동태학(Pharmacokinetics)이란 약물의 생체내에 있어서의 흡수, 분포, 비축, 대사, 배설의 과정을 연구하는 학문이다. 주요 변수 : Subject(실험대상), time(약투여시간 : 단위(hr)), conc(농도(concentration) : 단위(mcg/ml))	

데이터 셋	data.csv 데이터 셋에 관한 설명
data.csv 데이터 셋은 22개 관측치, 3개의 변수로 구성되어 있으며, 5명의 고객이 날짜별로 구매한 수량을 나타내고 있다. 날짜와 고객ID는 2개 이상의 중복 자료가 존재한다. 주요 변수 : Date(구매날짜), Customer(고객ID), Buy(구매수량)	

1) 척도(Scale)는 설문지와 같은 측정도구에서 응답자가 변인의 값을 선택할 수 있도록 일련의 기호 또는 숫자로 나타내어 변수를 측정하게 하는 단위이다. 척도에 관한 내용은 제9장에서 살펴본다.

데이터 셋	airquality 데이터 셋에 관한 설명

airquality 데이터 셋은 R에서 기본으로 제공되는 데이터 셋으로 New York의 대기에 대한 질을 측정한 데이터 셋이다. 전체 153개의 관측치와 6개의 변수로 구성되어 있으며, 변수명은 모두 대문자로 되어있다.

주요 변수 : Ozone(오존 수치),Solar.R(태양광), Wind(바람), Temp(온도),
Month(측정 월: 5~9), Day(측정 날짜 : 1~31일)

chap10_DescriptiveStatistics

데이터 셋	descriptive.csv 데이터 셋에 관한 설명

인구통계학적특성을 나타내는 변수를 기준으로 부모의 학력수준에 따라 자녀의 대학진학 합격여부를 조사한 데이터 셋으로 300개의 관측치와 8개의 변수로 구성되어 있다. 8개 변수(칼럼)에 대한 척도와 값의 범위는 다음과 같다.

resident	gender	age	level	cost	type	survey	pass
거주지역	성별	나이	학력수준	생활비	학교유형	만족도	합격여부
명목(1,2,3)	명목(1,2)	비율	서열(1,2,3)	비율	명목(1,2)	등간(5점)	명목(1,2)

chap11_CrossTableChi-squared

데이터 셋	diamonds 데이터 셋에 관한 설명

약 5만4천개의 다이아몬드에 관한 속성을 기록한 데이터 셋으로 53,940개의 관측치와 10개의 변수로 구성되어 있다. 주요 변수에 대한 설명은 다음과 같다.
price : 다이아몬드 가격($326~$18,823)
carat :다이아몬드 무게(0.2~5.01)
cut : 컷의 품질(Fair,Good,Very Good, Premium Ideal)
color : 색상(J:가장나쁨 ~ D:가장좋음)
clarity : 선명도(I1:가장나쁨, SI1, SI1, VS1, VS2, VVS1, VVS2, IF:가장좋음)
x: 길이 (0-10.74mm), y : 폭(0-58.9mm), z : 깊이 (0-31.8mm),
depth : 깊이 비율 = z / mean(x, y)

chap12_VisualizationAnalysis

데이터 셋	Chem97 데이터 셋에 관한 설명
mlmRev 패키지에서 제공되는 데이터 셋으로 1997년 영국 2,280개 학교 31,022명 학생을 대상으로 A레벨(대학시험) 화학점수를 기록한 데이터 셋이다. 전체 31,022개의 관측치와 8개의 변수로 구성되어 있다. 주요 변수 : lea(Local Education Authority) : 지방교육청(범위:1~15), school :학교 id(범위 : 1~132), student : 학생 id(범위 : 1~1250) score : A레벨 화학점수(범위:0,2,4,6,8,10), gender : 성별(범위:M, F), age : 18.5세 기준 월수(범위 : −6~+5), gcsescore : GCSE 개인평균성적(범위 : 0 ~ 8 사이 실수) ※ GCSE(General Certificate of Secondary Education)는 고등학교 재학 중에 치루는 수학능력인증시험을 의미한다.	

데이터 셋	VADeaths 데이터 셋에 관한 설명
R에서 기본으로 제공되는 데이터 셋으로 1940년 미국 버지니아주(Virginia)의 하위계층 사망비율을 기록한 데이터 셋이다. 전체 5행 4열의 numeric 자료형의 matrix 자료구조를 갖고 있다. 변수 구성 : Rural Male(시골출신 남자), Urban Male(도시출신 남자)) Rural Female(시골출신 여자), Urban Female(도시출신 여자))	

데이터 셋	quakes 데이터 셋에 관한 설명
R에서 제공하는 기본 데이터 셋으로 1964년 이후 피지(태평양)섬 근처에서 발생한 지진 사건에 관한 기록으로 전체 1,000개의 관측치와 5개의 변수로 구성되어 있다. 주요 변수 : lat(위도),long(경도),depth(수심:km),mag(리히터규모),stations(관측소)	

데이터 셋	SeatacWeather 데이터 셋에 관한 설명
latticeExtra 패키지에서 제공되는 데이터 셋으로 2007년 시애틀 타코마(Seattle-Tacoma) 공항의 1~3월 사이의 강수량과 온도가 기록되어 있으며, 전체 관측치 90개와 14개의 변수로 구성된 데이터프레임이다. 주요 변수 : month(1~3월), day(일), year(2007), max.temp(최고기온 : 화씨), record.max(최고기록온도), min.temp(최소온도), record.min(최소기록온도), precip(강수량 : 인치), record.precip(기록적인 강수량), time.max(최대온도시간), time.min(최소온도시간)	

데이터 셋	singer 데이터 셋에 관한 설명
lattice 패키지에서 제공되는 데이터 셋으로 'New York Choral Society' 합창단 성악가의 목소리 영역과 키 관계가 기록되어 있으며, 전체 관측치 235개와 2개의 변수로 구성된 데이터프레임이다. 주요 변수 : height(키 : 60~76), voice.part(목소리 영역 : 8개)	

데이터 셋	EuStockMarkets 데이터 셋에 관한 설명
datasets 패키지에서 제공되는 데이터 셋으로 1991~1998년 유럽의 주요 주식의 주가 지수 일일 마감 가격이 기록되어 있으며, 1,860 * 4의 matrix 자료구조가 트랜잭션(transaction) 자료구조로 변형된 시계열 자료이다. 1991년부터 1998년도에 의해서 4개의 칼럼(DAX(독일),SMI(스위스),CAC(프랑스),FTSE(영국))이 만들어지고, 각 칼럼 당 1,860개의 데이터로 구성되어 있다. 즉 년도별로 4개의 주식에 대한 주가 지수가 시계열 데이터로 표현된 데이터 셋이다.	

데이터 셋	USCancerRates 데이터 셋에 관한 설명
latticeExtra 패키지에서 제공되는 데이터 셋으로 1999 ~ 2003년도에 미국 도시에서 sex에 의한 암의 원인으로 사망한 비율이 기록되었으며, 전체 관측치 3,041개와 8개의 변수로 구성된 데이터프레임이다. 주요 변수 : rate.male(10만명 당 남자 사망률),LCL95.male(rate.male에 대한 95%신뢰 하한값),UCL95.male(rate.male에 대한 95%신뢰 상한값), rate.female (10만명 당 여자 사망률), state(미국 주), county(지도 경계)	

데이터 셋	mpg 데이터 셋에 관한 설명
ggplot2에서 제공하는 데이터 셋으로 1999년부터 2008년 사이의 가장 대중적인 모델 38개 자동차에 대한 연비효율을 기록한 데이터 셋으로 전체 관측 234개와 11개의 변수로 구성되어 있다. 주요 변수 : manufacturer(제조사), model(모델), displ(엔진크기), year(연식), cyl(실린더수), trans(변속기), drv(구동방식 : 사륜(4), 전륜(f), 후륜(r)), cty(gallon 당 도시 주행 마일수), hwy(gallon 당 고속도로 주행 마일수)	

데이터 셋	mtcars 데이터 셋에 관한 설명
ggplot2에서 제공하는 데이터 셋으로 자동차 모델에 관한 사양이 기록된 데이터 프레임이다. 전체 관측 32개와 11개의 변수로 구성되어 있다. 주요 변수 : mpg(연비), cyl(실린더 수), displ(엔진크기), hp(마력), wt(중량), qsec(1/4마일 소요시간), am(변속기:0=오토,1=수동), gear(앞쪽 기어 수), carb(카뷰레터 수)	

chap15_ClassificationAnalysis

데이터 셋	weather 데이터 셋에 관한 설명
날씨 관련 변수에 따라서 비가 내릴지의 여부를 기록한 데이터이다. 이 데이터를 분석하면 어떤 날씨 조건에 비가 내릴지 또는 내리지 않을지에 대한 판단 기준을 분석할 수 있다. 전체 관측치는 366개이고 15개의 변수로 구성되어 있다. 주요 변수 : Date(측정날짜),MinTemp(최저기온),MaxTemp(최대기온),Rainfall(강수량),Sunshine(햇빛),WindGustDir(돌풍방향),WindGustSpeed(돌풍속도),WindDir(바람방향),WindSpeed(바람속도),Humidity(습도),Pressure(기압),Cloud(구름),Temp(온도),RainToday(오늘 비 여부),RainTomorrow(내일 비 여부)	

chap17_AssociationAnalysis

데이터 셋	AdultUCI 데이터 셋에 관한 설명
arules패키지에서 제공되는 데이터 셋으로 성인을 대상으로 인구 소득에 관한 설문조사 데이터를 포함하고 있다. 전체 48,842개의 관측치와 15개 변수로 구성되어 있다. 주요 변수 : age(나이),workclass(직업:4개),education(교육수준:16개),marital-status(결혼상태:6개),occupation(직업:12개),relationship(관계:6개),race(인종:아시아계,백인),sex(성별),capital-gain(자본이득),capital-loss(자본손실),fnlwgt(미지의변수), hours-per-week(주당 근무시간), native-country(국가), income(소득)	

데이터 셋	Adult 데이터 셋에 관한 설명
arules패키지에서 제공되는 Adult는 성인을 대상으로 인구 소득에 관한 설문조사 데이터를 포함하고 있는 AdultUCI 데이터 셋을 트랜잭션 객체로 변환하여 준비된 데이터 셋이다. AdultUCI 데이터 셋은 전체 48,842개의 관측치와 15개 변수로 구성된 데이터 프레임이다.[2)] Adult 데이터 셋은 종속변수(Class)에 의해서 년 간 개인 수입이 $5만 이상 인지를 예측하는 데이터 셋으로 transactions 데이터로 읽어온 경우 48,842개의 transaction과 115개의 item으로 구성된다.	

데이터 셋	Groceries 데이터 셋에 관한 설명
arules패키지에서 제공되는 Groceries 데이터 셋은 1개월 동안 실제 로컬 식료품 매장에서 판매되는 트랜잭션 데이터를 포함하고 있다. 전체 9,835개의 트랜잭션(transaction)과 항목(item) 169 범주를 포함하고 있다.	

chap18_TimeseriesAnalysis

데이터 셋	WWWusage 데이터 셋에 관한 설명
R에서 제공되는 기본 데이터 셋으로 인터넷 사용 시간을 분 단위로 측정한 100개 시계열 데이터로 구성되어 있다.	

2) 제14장 지도학습의 AdultUCI 데이터 셋에 관한 설명 참고