빅데이터 분석

[R 데이터 탐색 - 다중변수 자료]

BigData Analysis

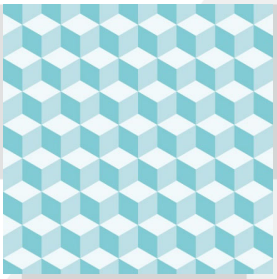
다중변수 자료탐색

데이터 탐색

산점도

상관분석

산점도



다중변수 자료탐색 - 산점도

산점도

- 산점도(scatter plot)이란 2개의 변수로 구성된 자료의 분포를 알아보는 그래프
- 관측단위별 값들의 분포를 통해서 2개의 변수 사이의 관계를 파악

변수 -> 데이터프레임의 열

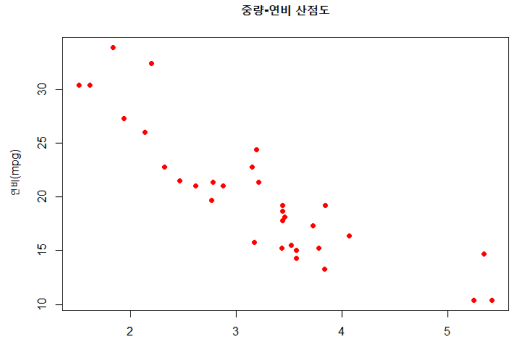
	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225.0	105	2.76	3.460	20.22	1	0	3	1
Duster 360	14.3	8	360.0	245	3.21	3.570	15.84	0	0	3	4
Merc 240D	24.4	4	146.7	62	3.69	3.190	20.00	1	0	4	2
Merc 230	22.8	4	140.8	95	3.92	3.150	22.90	1	0	4	2
Merc 280	19.2	6	167.6	123	3.92	3.440	18.30	1	0	4	4
Merc 280C	17.8	6	167.6	123	3.92	3.440	18.90	1	0	4	4

관측단위 -> 데이터프레임의 행

다중변수 자료탐색 – 산점도

산점도

- 산점도(scatter plot)이란 2개의 변수로 구성된 자료의 분포를 알아보는 그래프
- 관측단위별 값들의 분포를 통해서 2개의 변수 사이의 관계를 파악



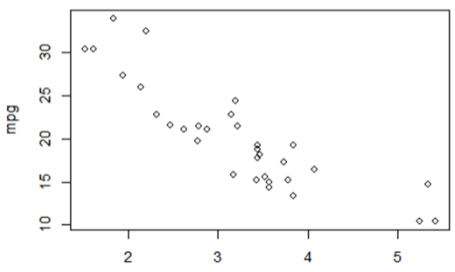
변수 -> x축의 “중량”
변수 -> y축의 “연비”

변수 “중량”과 “연비 " 가 관측단위별로
어떻게 분포되어있는지 알 수 있습니다.
변수 “중량 " 에 따라서 “변수 " 연비 " 가
어떻게 달라지는지 파악할 수 있습니다.

다중변수 자료탐색 – 산점도

산점도 그리기

```
print(mtcars)
wt <- mtcars$wt
mpg <- mtcars$mpg
plot(wt, mpg)
```



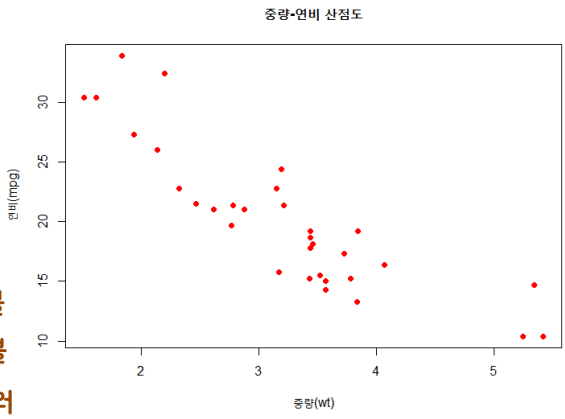
mtcars dataset

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3
Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3
Valiant	18.1	6	225.0	105	2.76	3.460	20.22	1	0	3
Duster 360	14.3	8	360.0	245	3.21	3.570	15.84	0	0	3
Merc 240D	24.4	4	146.7	62	3.69	3.190	20.00	1	0	4
Merc 230	22.8	4	140.8	95	3.92	3.150	22.90	1	0	4
Merc 280	19.2	6	167.6	123	3.92	3.440	18.30	1	0	4
Merc 280C	17.8	6	167.6	123	3.92	3.440	18.90	1	0	4
Merc 450SE	16.4	8	275.8	180	3.07	4.070	17.40	0	0	3
Merc 450SL	17.3	8	275.8	180	3.07	3.730	17.60	0	0	3
Merc 450SLC	15.2	8	275.8	180	3.07	3.780	18.00	0	0	3
Cadillac Fleetwood	10.4	8	472.0	205	2.93	5.250	17.98	0	0	3
Lincoln Continental	10.4	8	460.0	215	3.00	5.424	17.82	0	0	3
Chrysler Imperial	14.7	8	440.0	230	3.23	5.345	17.42	0	0	3
Fiat 128	32.4	4	78.7	66	4.08	2.200	19.47	1	1	4
Honda Civic	30.4	4	75.7	52	4.93	1.615	18.52	1	1	4
Toyota Corolla	33.9	4	71.1	65	4.22	1.835	19.90	1	1	4
Toyota Corona	21.5	4	120.1	97	3.70	2.465	20.01	1	0	3
Dodge Challenger	15.5	8	318.0	150	2.76	3.520	16.87	0	0	3
AMC Javelin	15.2	8	304.0	150	3.15	3.435	17.30	0	0	3
Camaro Z28	13.3	8	350.0	245	3.73	3.840	15.41	0	0	3
Pontiac Firebird	19.2	8	400.0	175	3.08	3.845	17.05	0	0	3
Fiat X1-9	27.3	4	79.0	66	4.08	1.935	18.90	1	1	4
Porsche 914-2	26.0	4	120.3	91	4.43	2.140	16.70	0	1	5
Lotus Europa	30.4	4	95.1	113	3.77	1.513	16.90	1	1	5
Ford Pantera L	15.8	8	351.0	264	4.22	3.170	14.50	0	1	5
Ferrari Dino	19.7	6	145.0	175	3.62	2.770	15.50	0	1	5
Maserati Bora	15.0	8	301.0	335	3.54	3.570	14.60	0	1	5
Volvo 142E	21.4	4	121.0	109	4.11	2.780	18.60	1	1	4

다중변수 자료탐색 – 산점도

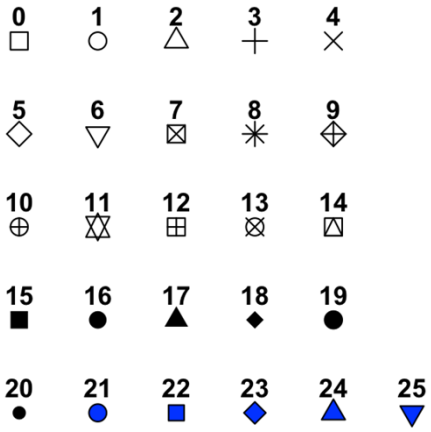
산점도 그리기

```
data(mtcars)          # R 제공 mtcars 데이터셋
wt <- mtcars$wt       # 중량 자료
mpg <- mtcars$mpg     # 연비 자료
plot(wt, mpg,         # 2개 변수 (x축, y축)
     main = "중량-연비 산점도", # 제목
     xlab = "중량(wt)",       # x축 레이블
     ylab = "연비(mpg)",      # y축 레이블
     col = "red",             # point 컬러
     pch = 19)                # point 종류
```



다중변수 자료탐색 – 산점도

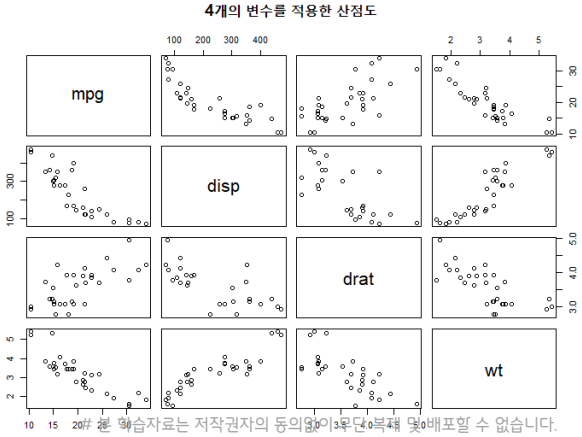
산점도 pch(plot characters)값에 따른 점의 모양



다중변수 자료탐색 – 산점도

3개 이상의 변수 사이의 산점도

- 3개 이상의 변수 사이의 관계를 파악하고 싶은 경우, 여러 개의 산점도 그리기 가능



다중변수 자료탐색 – 산점도

3개 이상의 변수 사이의 산점도

pairs( ) 함수

- 여러 개의 변수에 대해 짝지어진 산점도를 한번에 그리는 함수
- pairs(target, main = " Multi Plots")

대괄호[ ]

- 데이터프레임 형식은 대괄호[ ]를 사용하여 데이터를 조회, 추출

다중변수 자료탐색 – 산점도

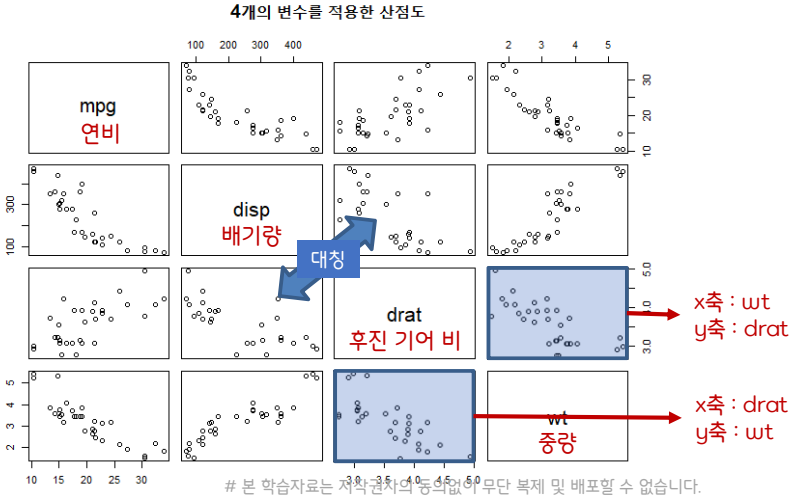
R 제공 mtcar 데이터셋 이용 다중 산점도

```
vars <- c("mpg", "disp", "drat", "wt")  # 대상 변수
target <- mtcars[,vars]  # 데이터프레임에서 위의 변수명을 가진 열 선택
head(target)  # 데이터프레임 확인
pairs(target, main =  " Multi Plots") # 다중 산점도 생성
```

head(데이터프레임, 숫자)
데이터가 너무 클 때, 해당 숫자까지만 보여줌

다중변수 자료탐색 – 산점도

3개 이상의 변수 사이의 산점도



다중변수 자료탐색 – 산점도

실습

상명이는 미국 주별 도심 인구(UrbanPop), 폭력 사건 수(Assault), 살인 사건 수(Murder) 변수들 사이의 상관관계를 보여주는 산점도를 시각화 해보고자 합니다.

R 제공 USArrests 데이터셋을 사용하여 여러가지 변수들 사이의 상관관계를 보여주는 산점도를 만들고, x축에 주별 도심 인구수(UrbanPop) 및 y축에 폭력 사건 수(Assault)가 있는 산점도를 찾아서 표시하세요.

** 산점도 표시할 때, 캡처 도구 내 펜을 활용하여 표시하면 편리합니다.*

다중변수 자료탐색 – 산점도

3개 이상의 변수 사이의 산점도

먼저 데이터셋 확인

Print(USArrests)

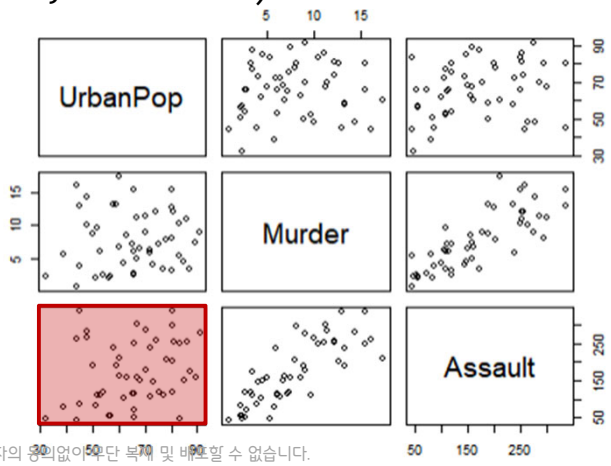
USArrests dataset

	Murder	Assault	UrbanPop	Rape
Alabama	13.2	236	58	21.2
Alaska	10.0	263	48	44.5
Arizona	8.1	294	80	31.0
Arkansas	8.8	190	50	19.5
California	9.0	276	91	40.6
Colorado	7.9	204	78	38.7
Connecticut	3.3	110	77	11.1
Delaware	5.9	238	72	15.8
Florida	15.4	335	80	31.9
Georgia	17.4	211	60	25.8
Hawaii	5.3	46	83	20.2
Idaho	2.6	120	54	14.2
Illinois	10.4	249	83	24.0
Indiana	7.2	113	65	21.0
Iowa	2.2	56	57	11.3
Kansas	6.0	115	66	18.0
Kentucky	9.7	109	52	16.3
Louisiana	15.4	249	66	22.2
Maine	2.1	83	51	7.8
Maryland	11.3	300	67	27.8
Massachusetts	4.4	149	85	16.3
Michigan	12.1	255	74	35.1
Minnesota	2.7	72	66	14.9
Mississippi	16.1	259	44	17.1
Missouri	9.0	178	70	28.2
Montana	6.0	109	53	16.4
Nebraska	4.3	102	62	16.5
Nevada	12.2	252	81	46.0
New Hampshire	2.1	57	56	9.5
New Jersey	7.4	159	89	18.8
New Mexico	11.4	285	70	32.1
New York	11.1	254	86	26.1
North Carolina	13.0	337	45	16.1
North Dakota	0.8	45	44	7.3
Ohio	7.3	120	75	21.4
Oklahoma	6.6	151	68	20.0
Oregon	4.9	159	67	29.3
Pennsylvania	6.3	106	72	14.9
Rhode Island	3.4	174	87	8.3
South Carolina	14.4	279	48	22.5
South Dakota	3.8	86	45	12.8
Tennessee	13.2	188	59	26.9
Texas	12.7	201	80	25.5
Utah	3.2	120	80	22.9
Vermont	2.2	48	32	11.2
Virginia	8.5	156	63	20.7
Washington	4.0	145	73	26.2
West Virginia	5.7	81	39	9.3
Wisconsin	2.6	53	66	10.8
Wyoming	6.8	161	60	15.6

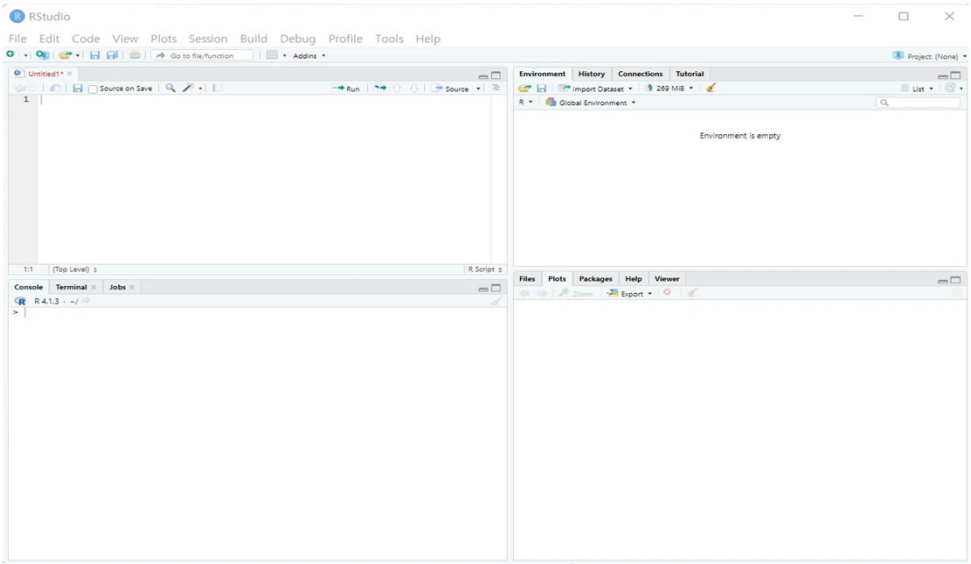
다중변수 자료탐색 – 산점도

3개 이상의 변수 사이의 산점도

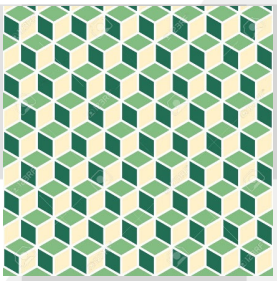
```
vars <- c("UrbanPop", "Murder", "Assault")
target <- USArrests[,vars]
pairs(target)
```



다중변수 자료탐색 – 산점도



상관분석



BigData Analysis

다중변수 자료탐색 - 상관분석

상관분석과 상관계수

- 상관분석(correlation analysis) 은 두 변수간의 관계를 분석하기 위해 사용
- 변수는 연속형 자료만 가능(구간척도, 비율척도)

- 예시. 키의 변화는 몸무게의 변화와 관계가 있는지
  - 가설: 키가 커지면 몸무게가 늘어난다.

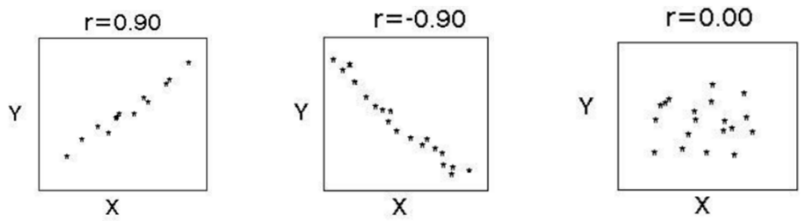
데이터 구조 - 변수 개수 기반 분류
- 단일변수 자료
- 다중변수 자료

데이터 구조 - 자료 특성 기반 분류
- 범주형 자료 - 명목척도 / 서열척도
- 연속형 자료 - 구간 척도 / 비율 척도

다중변수 자료탐색 – 상관분석

상관분석과 상관계수

- 상관계수(correlation coefficient) r = **X와 Y가 함께 변하는 정도** / X와 Y가 각각 변하는 정도
  - 두 변수간 X와 Y가 완전히 동일하면 상관계수 r은 +1 (양의상관관계)
  - 두 변수간 X와 Y가 반대방향으로 완전히 동일하면 상관계수 r은 -1 (음의 상관관계)
  - 두 변수간 X와 Y가 상관성이 없으면 상관계수 r은 0



다중변수 자료탐색 – 상관분석

상관분석과 상관계수

beers	5	2	9	8	3	7	3	5	3	5
bal	0.10	0.03	0.19	0.12	0.04	0.095	0.07	0.06	0.02	0.05

- 음주정도와 혈중알코올농도가 상관성이 있는지 알아보는 예시
  - 10명의 실험자들에 대해 맥주를 마신 잔수(beers)와 혈중알코올농도(bal)에 대한 측정자료

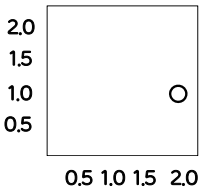
*bal : blood alcohol concentration

  - 음주정도에 따라 혈중알코올농도가 변하는 정도

다중변수 자료탐색 - 상관분석

plot() 함수

- plot() 함수는 x와 y의 2개 축을 기준으로 좌표를 찍듯이 그리는 컨셉을 가지는 함수
  - 예시. plot(2,1)



- plot() 함수는 산점도(scatter plot)를 그리는 함수
- plot(벡터2(Y)~ 벡터1(X), data=데이터프레임)
  - 벡터1과 벡터2의 관계 → 벡터1(X)이 변화함에 따라 벡터2(Y)가 변화하는 정도

다중변수 자료탐색 - 상관분석

plot() 함수

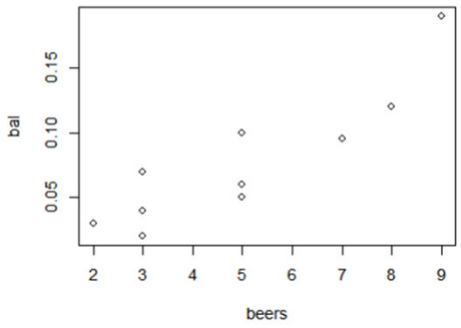
- 데이터프레임을 생성하고 산점도
- 음주정도(beers)에 따라 혈중알코올농도(bal)가 변하는 정도

```
beers <- c(5, 2, 9, 8, 3, 7, 3, 5, 3, 5)
bal <- c(0.1, 0.03, 0.19, 0.12, 0.04, 0.095, 0.07, 0.06, 0.02, 0.05)
ca <- data.frame(beers, bal)
print(ca)
plot(bal~beers, data=ca) # 산점도
```

다중변수 자료탐색 – 상관분석

plot() 함수

```
> beers <- c(5, 2, 9, 8, 3, 7, 3, 5, 3, 5)
> bal <- c(0.1, 0.03, 0.19, 0.12, 0.04, 0.095, 0.07, 0.06, 0.02, 0.05)
> ca <- data.frame(beers, bal)
> print(ca)
   beers   bal
1      5 0.100
2      2 0.030
3      9 0.190
4      8 0.120
5      3 0.040
6      7 0.095
7      3 0.070
8      5 0.060
9      3 0.020
10     5 0.050
> plot(bal~beers, data=ca)
>
```



다중변수 자료탐색 – 상관분석

lm() 함수

- lm() 함수는 linear model 약자로 선형모델을 맞추는데 사용
- lm() 함수는 두 변수의 선형관계를 가장 잘 나타낼 수 있는 선의 식을 자동으로 찾는 역할
- lm() 함수는 'y=ax+b' 형태의 1차식
- abline() 함수 – 그래프 위에 선을 추가하고 싶은 경우

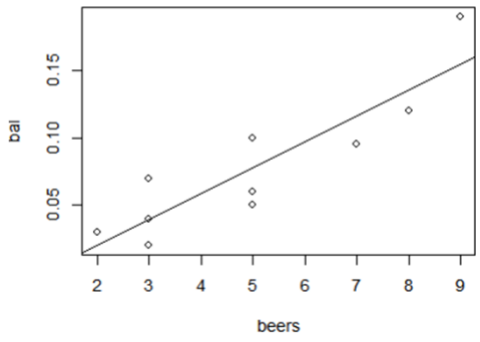
- lm(벡터2~벡터1, data=데이터프레임) # 회귀식 도출
- abline(회귀식) # 회귀선 그리기

```
camodel <- lm(bal~beers, data=ca)
abline(camodel)
```

다중변수 자료탐색 – 상관분석

lm() 함수

```
> plot(bal~beers, data=ca)
> camodel <- lm(bal~beers, data=ca)
> abline(camodel)
> |
```



다중변수 자료탐색 – 상관분석

cor() 함수

- 상관계수(correlation)를 구하는 함수
- method를 “person”, “kendall”,”spearman”으로 지정할 수 있는데, 기본값은 “person”(피어슨 상관계수)

- cor(벡터1,벡터2) # 상관계수 계산

cor(beers,bal)

피어슨 상관계수

$$\text{피어슨상관계수} = \frac{\text{공분산}}{\text{표준편차} \cdot \text{표준편차}}$$

편차 = 평균과 예측값 간의 차이 + 예측값과 실제값의 차이
공분산은 두 개의 확률 변수의 선형관계를 나타내는 값이다.

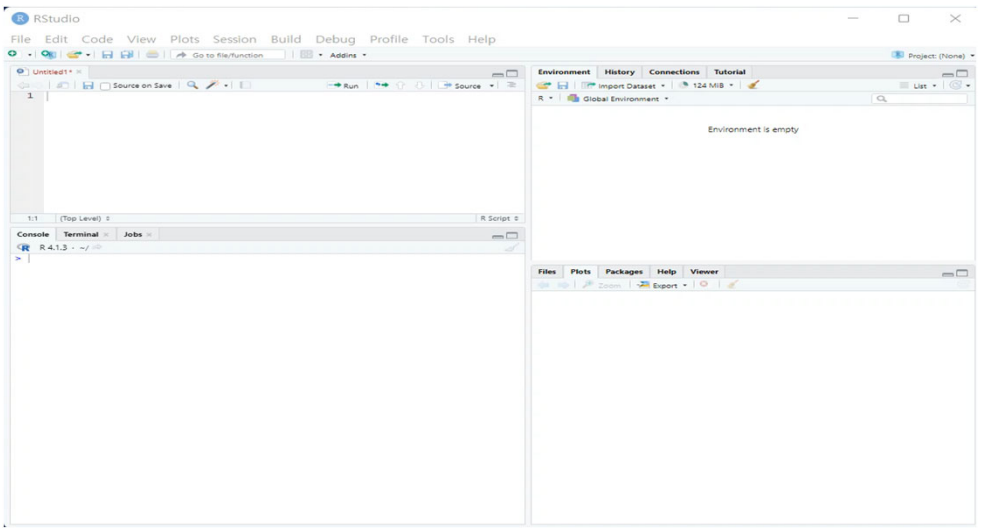
다중변수 자료탐색 - 상관분석

cor() 함수

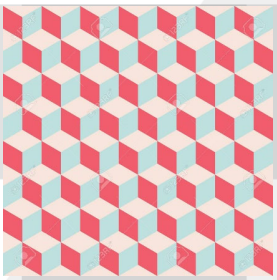
```
> cor(beers,bal)
[1] 0.8882323
> |
```

상관계수	상관관계
± 0.9 이상	상관관계가 아주 높다
± 0.7 ~ 0.9	상관관계가 높다
± 0.4 ~ 0.7	상관관계가 있다
± 0.2 ~ 0.4	상관관계가 있으나 낮다
± 0.2 미만	상관관계가 거의 없다

다중변수 자료탐색 - 상관분석



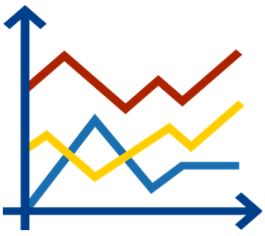
선그래프



다중변수 자료탐색 - 선그래프

R 선그래프

- 두개의 변수 중 하나가 시간을 나타내는 값일 때 사용
- 시계열 자료(times series data) - 시간의 변화에 따라 자료의 증감추이를 확인



다중변수 자료탐색 – 선그래프

R 선그래프

month	1	2	3	4	5	6	7	8	9	10	11	12
cold	5	8	7	9	4	6	12	13	8	6	6	4

- 한개 학교의 월별 감기 환자 통계를 알아본 예시

다중변수 자료탐색 – 선그래프

plot() 함수

- 선그래프를 작성하는 함수는 산점도를 작성할 때 사용한 plot()함수
- plot() 함수에서 매개변수 type의 값을 “l” (알파벳)로 하면 선그래프가 작성
- plot(x축 데이터, y축 데이터, 옵션)

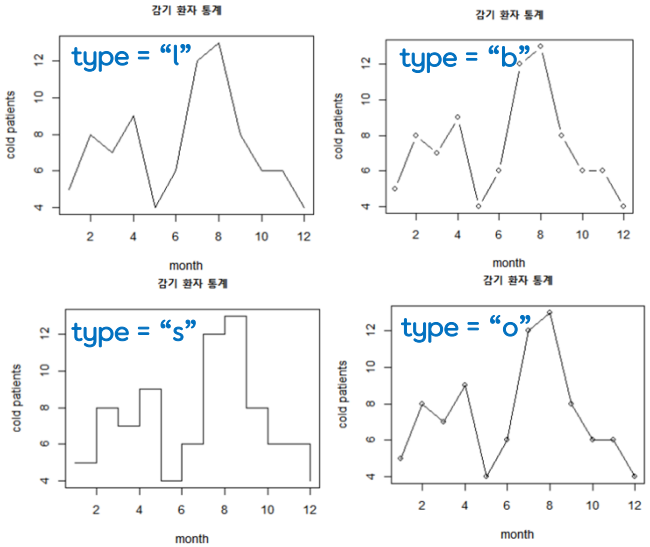
```
month <- 1:12
cold <- c(5,8,7,9,4,6,12,13,8,6,6,4)
plot(month,                          # x data
     cold,                           # y data
     main="감기 환자 통계",           # 제목
     type="l",                       # 그래프의 종류 선택(알파벳) Line
     lty=1,                          # 선의 종류(Line Type) 선택
     lwd=1,                          # 선의 굵기 선택
     xlab="month",                   # x축 레이블
     ylab="cold patients")           # y축 레이블
```

다중변수 자료탐색 – 선그래프

plot() 함수

```
> month <- 1:12
> cold<- c(5,8,7,9,4,6,12,13,8,6,6,4)
> plot(month,
+      cold,
+      main="감기 환자 통계",
+      type="l",
+      lty=1,
+      lwd=1,
+      xlab="month",
+      ylab="cold patients"
+      )
>
```

lty (line type)



다중변수 자료탐색 – 선그래프

R 복수 선그래프

month	1	2	3	4	5	6	7	8	9	10	11	12
cold1	5	8	7	9	4	6	12	13	8	6	6	4
cold2	4	6	5	8	7	8	10	11	6	5	7	3

- 두개 학교의 월별 감기 환자 통계를 알아본 예시

다중변수 자료탐색 – 선그래프

lines() 함수

- lines()함수는 좌표의 점들을 이어서 선을 그리는 함수
- plot() 함수로 작성한 그래프 위에 선을 겹쳐서 그리는 역할

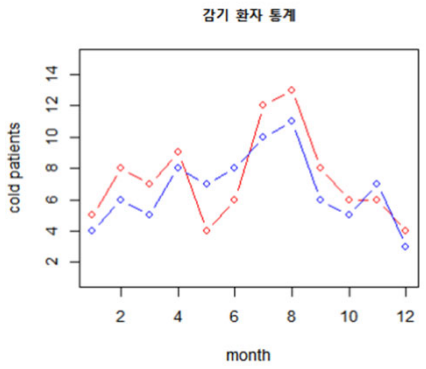
```
month <- 1:12
cold1 <- c(5,8,7,9,4,6,12,13,8,6,6,4)
cold2 <- c(4,6,5,8,7,8,10,11,6,5,7,3)
plot(month,                  # x data
     cold1,                  # y data
     main="감기 환자 통계",  # 제목
     type="b",               # 그래프의 종류 선택(알파벳) Line
     lty=1,                  # 선의 종류(Line Type) 선택
     lwd=1,                  # 선의 굵기 선택
     col="red",              # 선의 색 선택(빨강)
     xlab="month",           # x축 레이블
     ylab="cold patients"    # y축 레이블
     ylim=c(1,15)            # y축 값의 (하한, 상한)
)
```

```
lines(month,      # x data
      cold2,      # y data
      type="b",   # 선의 종류 선택
      col="blue") # 선의 색 선택(파랑)
```

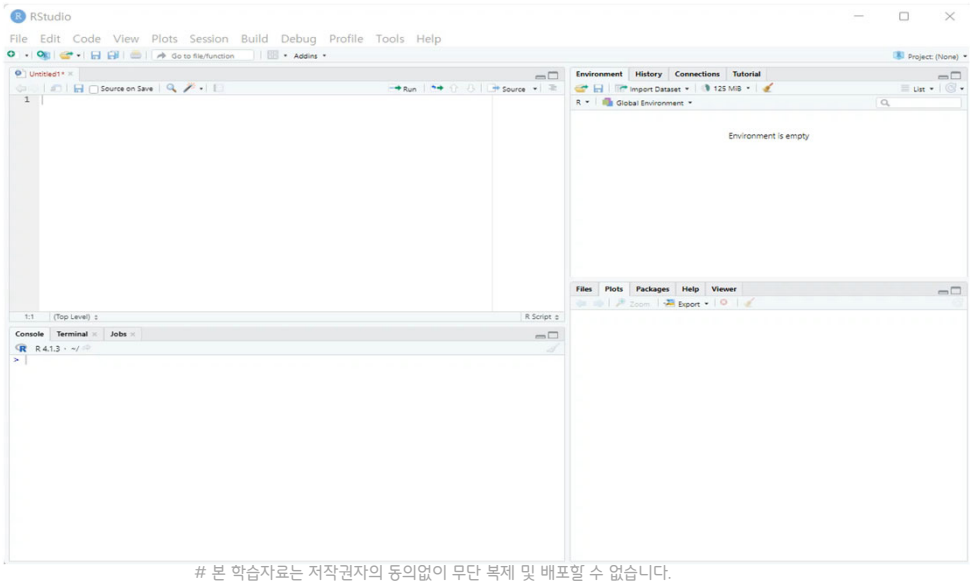

다중변수 자료탐색 – 선그래프

lines() 함수

```
> month <- 1:12
> cold1<- c(5,8,7,9,4,6,12,13,8,6,6,4)
> cold2<- c(4,6,5,8,7,8,10,11,6,5,7,3)
> plot(month,
+      cold1,
+      main="감기 환자 통계",
+      type="b",
+      lty=1,
+      lwd=1,
+      col="red",
+      xlab="month",
+      ylab="cold patients",
+      ylim=c(1,15)
+ )
>
> lines(month,
+       cold2,
+       type="b",
+       col="blue")
>
```



다중변수 자료탐색 - 선그래프



SUMMARY

- **산점도**
  - plot(x축, y축)
  - main / xlab / ylab / col / pch
- **다중산점도**
  - pairs(target, main = " Multi Plots")
  - target <- 데이터프레임[,vars]

SUMMARY

- **상관분석**
  - 연속형 자료로만 가능
  - plot() 함수 # 산점도
  - lm(벡터2~벡터1, data=데이터프레임) # 회귀식 도출
  - abline(회귀식) # 회귀선 그리기
- **상관계수**
  - r = X와 Y가 함께 변하는 정도
  - cor(벡터1,벡터2) # 상관계수 계산
  - r은 +1 (양의상관관계) / R은 -1 (음의 상관관계) / R은 0

SUMMARY

- **선그래프**
  - plot() 함수 # 산점도
  - lines() 함수 # plot() 함수로 작성한 그래프 위에 선을 겹쳐서 그리는 역할

감사합니다.