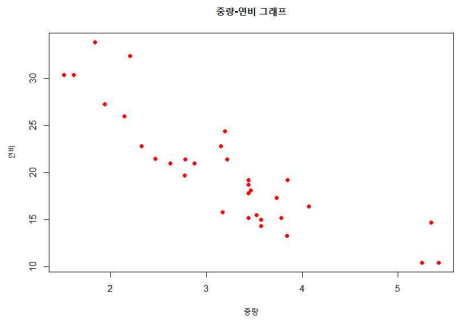
다중변수 자료 : 변수가 2개 이상인 자료(2차원). 변수는 열로, 개별 관측값들이 행으로 이루어진다.

산점도 : 2개의 변수로 구성된 자료의 분포를 알아보는 그래프. 관측값들의 분포를 통해 2개의 변수 사이의 관계 파악. **plot()**

```
wt <- mtcars$wt                    #자동차 중량
mpg <- mtcars$mpg                  #자동차 연비
plot(wt, mpg,                      #2개의 변수
     main="중량-연비 그래프", #제목
     xlab="중량",                  #x축 레이블
     ylab="연비",                  #y축 레이블
     col="red",                    #point color
     pch=19)                       #point 종류(점 모양 변경)
```

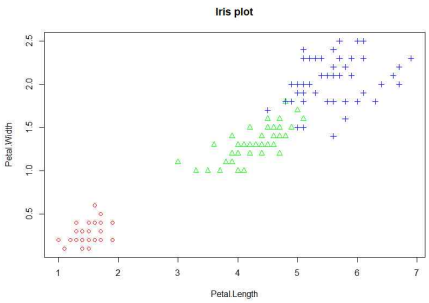


아래 세 방식중 원하는걸로 골라 쓰면 된다.

plot(mtcars$wt, mtcars$mpg) / plot(mtcars[,c("wt","mpg")] / plot(mpg~wt, data=mtcars)

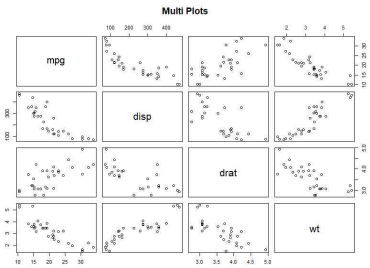
그룹 정보가 있을 때 :

```
#iris 데이터셋에서 3~4열 값들만 추출
iris.2 <- iris[,3:4]
#팩터 타입으로 된 iris$Species를 1,2,3으로 변경
point <- as.numeric(iris$Species)
point
#점의 색 설정
color <- c("red","green","blue")
plot(iris.2, main="Iris plot", pch=c(point),col=color[point])
```



다중산점도 : 여러 개의 변수를 짝지어 한 번에 산점도를 그린 것. **pairs()**

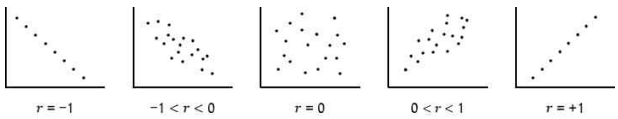
```
vars<-c("mpg","disp","drat","wt")   #대상 변수
target <- mtcars[,vars]
head(target)                        #각 변수 6개 출력
pairs(target, main="Multi Plots")   #다중산점도
```



상관분석 : 두 변수가 어느 정도의 선형적 관계가 있는지를 시각적 방법이 아닌 수치상으로 나타낼 수 있는 방법.

피어슨 상관계수 : $r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$ 여기서 r이 피어슨 상관계수 (상관계수라 줄여서 씀). 선형선의 정도를 나타내는 척도임. 항상 -1 <= r <= 1 이다. r이 -1이나 1에 가까워질수록 분포가 직선에 가까워지고, 0에 가까워질수록 두 변수간에 상관관계가 없다. cor()함수를 사용해 구한다.

r > 0 인 경우 : x가 증가하면 y도 증가. 양의 상관 관계
r < 0 인 경우 : x가 증가하면 y는 감소. 음의 상관 관계



lm()함수를 이용하면 두 변수의 선형 관계를 가장 잘 나타낼 수 있는 선의 식(회귀식)을 찾아준다.
abline()함수를 사용하면 회귀식을 이용하여 산점도 위에 회귀선을 그려준다

시계열 자료 : 시간의 변화에 따라 수집한 자료. 다양한 분석 방법이 존재하고, 대표적으로 선 그래프가 있다.
선 그래프는 plot()함수의 매개변수를 type="l"로 바꾸면 된다.

요약

01 초기에 확보한 데이터를 정제하고 가공하여 분석에 적합한 데이터를 얻는 과정을 **데이터 전처리**data preprocessing라고 한다.

02 **결측값**missing value은 데이터를 수집하고 저장하는 과정에서 저장할 값을 얻지 못하는 경우 발생한다.

03 결측값을 처리하는 방법은 다음의 두 가지 방법이 있다.

- 결측값을 제거하거나 제외시킨 다음 데이터를 분석한다.
- 결측값을 추정하여 적당한 값으로 치환한 후 데이터를 분석한다.

04 R에서는 결측값을 표현하기 위해 NA라고 하는 특별한 값을 제공하는데, NA는 숫자형, 문자형, 논리형 데이터 어디에서나 결측값을 나타내는 용도로 사용할 수 있다.

05 **특이값**outlier은 정상적이라고 생각되는 데이터의 분포 범위 밖에 위치하는 값들을 말하며, '이상치'라고도 부른다.

06 데이터셋에 특이값이 포함되어 있는지의 여부는 다음과 같은 기준을 가지고 찾는다.

① 논리적으로 있을 수 없는 값이 있는지 찾아본다.

② 상식을 벗어난 값이 있는지 찾아본다.

③ 상자그림boxplot을 통해 찾아본다.

07 **정렬**sort은 데이터를 주어진 기준에 따라 크기순으로 재배열하는 과정을 말한다.

08 데이터셋의 열 값을 기준으로 여러 개의 데이터셋으로 분리할 때는 `split()` 함수를 이용하고, 데이터셋으로부터 조건에 맞는 행row들을 추출할 때는 `subset()` 함수를 이용한다.

09 **샘플링**sampling은 주어진 값들이 있을 때 그중에서 임의의 개수의 값들을 추출하는 작업을 의미한다.

10 **조합**combination은 주어진 데이터값 중에서 몇 개씩 짝을 지어 추출하는 작업을 말한다. R에서는 `combn()` 함수를 사용하는데, 결과에서 각 열이 하나의 조합을 의미한다.

11 2차원 데이터셋에서 데이터 그룹에 대해 합계나 평균 등을 계산해야 하는 작업을 **집계**aggregation라고 하며 R에서는 `aggregate()` 함수를 통해서 작업 가능하다.

12 여러 파일에 흩어져 있는 자료를 공통 열을 연결 고리로 하여 하나로 합치는 작업을 **병합**merge이라고 한다.

■ 함수 정리

전처리 관련 함수	
`is.na()`	벡터 값들에 대해 결측값 여부를 확인한다.
`complete.cases()`	데이터프레임에서 결측값이 없는 행들의 번호를 반환한다.
`boxplot.stats()$out`	벡터에 포함된 특이값을 반환한다.
`order()`	벡터 값들이 크기순으로 몇 번째인지를 알려준다.
`sort()`	벡터 값들을 크기순으로 정렬한다.
`split()`	벡터, 데이터프레임을 기준(열)의 값에 의해 분리한다.
`subset()`	데이터프레임에서 기준에 맞는 행들만 추출한다.
`sample()`	주어진 값들에서 임의의 개수의 값을 추출한다.
`set.seed()`	`sample()` 함수가 일정한 결과를 도출하도록 설정한다.
`combn()`	주어진 값들에서 n개의 값을 짝지어 추출한다.
`aggregate()`	2차원 데이터를 그룹값에 의해 집계한다.
`merge()`	2개의 데이터프레임을 공통 열에 의해 연결하여 하나의 데이터프레임을 생성한다.

요약

01 데이터 시각화data visualization는 숫자 형태의 데이터를 그래프나 그림 등의 형태로 표현하는 것을 말한다.

02 트리맵tree map은 사각 타일의 형태로 구성되어 있으며, 각 타일의 크기와 색으로 데이터의 크기를 나타낸다. R에서 트리맵을 작성하기 위해서는 먼저 `treemap` 패키지를 설치해야 한다.

03 버블 차트bubble chart는 산점도 위에 버블의 크기로 정보를 표시하는 시각화 방법이다. 산점도가 2개의 변수에 의한 위치 정보를 표시한다면 버블 차트는 3개의 변수 정보를 하나의 그래프에 표시한다.

04 모자이크 플롯mosaic plot은 다중변수 범주형 데이터에 대해 각 변수의 그룹별 비율을 면적으로 표시하여 정보를 전달한다.

05 `ggplot`은 R에서 미적인 그래프를 작성할 때 널리 활용된다.

06 `ggplot` 명령문은 여러 개의 함수들을 연결하여 사용한다. 일반적인 `ggplot` 명령문의 형태는 다음과 같다.

```
ggplot(data=xx, aes(x=x1,y=x2)) +
  geom_xx() +
  geom_yy() +
  ...
```

07 차원 축소dimension reduction란 고차원 데이터를 2, 3차원 데이터로 축소하는 기법으로 고차원상의 데이터 분포를 2, 3차원상에서 관찰할 수 있다.

■ 함수 정리

데이터 시각화 관련 함수	
treemap()	트리맵을 작성한다.
symbols()	버블 차트를 작성한다.
text()	버블 차트 위에 텍스트를 출력한다.
mosaicplot()	모자이크 플롯을 작성한다.
ggplot()	ggplot 작성에 필요한 기본 정보를 지정한다.
geom_bar()	ggplot에서 막대그래프를 작성한다.
ggtitle()	ggplot에서 그래프의 제목을 설정한다.
geom_histogram()	ggplot에서 히스토그램을 작성한다.
geom_point()	ggplot에서 산점도를 작성한다.
geom_boxplot()	ggplot에서 상자그림을 작성한다.
geom_line()	ggplot에서 선그래프를 작성한다.
Rtsne()	고차원 데이터를 2, 3차원 데이터로 축소한다.
scatter3d()	Rtsne() 함수의 결과를 이용하여 3차원 산점도를 작성한다.

요약

01 R에서 구글맵을 사용하기 위해서는 다음과 같은 준비 절차가 있다.

① R을 최신 버전으로 설치한다.

② ggplot2를 최신 버전으로 업데이트한다.

③ ggmap 패키지를 설치한다.

④ 구글맵을 사용하기 위한 API 키를 얻는다.

02 R에서는 지도 위에 마커와 텍스트를 표시할 수 있다. **마커** marker란 지도상에서 특정 지점의 위치에 표시하는 기호이다.

03 구글맵 위에는 마커나 텍스트뿐만 아니라 ggplot 패키지를 이용하여 원과 같은 도형도 표시할 수 있다.

■ 함수 정리

지도 관련 함수	
register_google()	구글맵 이용을 위한 API 키를 입력한다.
geocode()	특정 지점의 위도와 경도 값을 반환한다.
get_googlemap()	특정 지점 근방의 지도를 가져온다.
ggmap()	가져온 지도를 화면에 출력한다.
geom_text()	지도 위에 텍스트를 표시한다.
geom_point()	지도 위에 원을 표시한다.

11장

독립변수 : 어떤 현상을 설명할 때 현상의 발생에 영향을 미치는 요인들(=설명변수 = x)
종속변수 : 영향에 따라 값이 결정되는 요인들 (=반응변수 = y)
예측모델 : 독립변수에 해당하는 자료와 종속변수에 해당하는 자료를 모아 관계를 분석하고 이를 예측에 사용할 수 있는 통계적 방법으로 정리한 것
회귀분석 : regression analysis
회귀 이론을 기초로 독립변수가 종속변수에 미치는 영향을 파악하여 예측모델(회귀식)을 도출하는 통계적 방법
lm()함수로 회귀식을 구하면 된다. abline()함수로 시각화 #11-1
독립변수 한 개 : 단순 회귀 #11-1~3
독립변수 두 개 이상 : 다중 회귀 (예시로 키,몸무게가 x이고, 혈당수치가 y인 경우)
stepAIC()함수로 효율적인 다중선형 회귀모델을 위한 변수를 선별한다. #11-4~5

분류 : 데이터로부터 어떤 범주를 예측하는 작업. 이 문제를 회귀의 방법으로 푸는게 로지스틱 회귀
로지스틱 회귀 : logistic regression
회귀모델에서 종속변수의 값의 형태가 연속형 숫자가 아닌 범주형 값인 경우를 다루기 위해서 만들어진 통계적 방법
종속변수가 숫자로 표현되어야 한다.
일반 회귀와 다르게 **glm()함수**로 로지스틱 회귀모델을 구한다. #11-6~8

12장

머신러닝 : 방대한 데이터를 컴퓨터가 스스로 분석하고 학습하여 유용한 정보를 얻어내거나
미래를 예측하기 위한 예측모델을 만들어내는 기술

군집화 : clustering. 주어진 대상의 데이터들을 유사성이 높은 것끼리 묶어주는 기술.
이런 묶음을 군집, 범주, 그룹 등 다양한 용어로 부름
비지도학습에 해당한다.

k-평균 : 먼저 군집의 중심점을 잡고, 다른 점들을 거리가 가장 가까운 중심점의 군집에 속하는 것으로 결정하는 군집화 분류 방법.
군집의 중심점을 잡을 때, 각 점들의 평균을 이용한다.
kmeans()함수로 군집화 실행하고, clusplot()함수로 시각화한다. #12-1
모든 변수가 거리 계산에 동등한 영향을 갖도록 하기 위해서 자료의 범위를 0~1 사이로 표준화해야 함. #12-2

분류 : classification. 그룹의 형태로 알려진 데이터들이 있을 때 그룹을 모르는 어떤 데이터에 대해 어느 그룹에 속하는지 예측.
지도학습에 해당한다.

k-최근접 이웃 : 1. 그룹을 모르는 데이터 P에 대해 이미 그룹이 알려진 데이터 중 P와 가장 가까이에 있는 k개의 데이터를 수집
2. k개의 데이터가 가장 많이 속해 있는 군집을 P의 군집으로 결정.
knn()함수 사용한다. #12-3

k-fold 교차 검증 : 데이터를 임의로 훈련용과 테스트용으로 나누어 모델을 개발하는 과정을 여러 번 반복하여 그곳에서 도출되는 예측 정확도의 평균을 구하는 것을 체계화한 방법론. cvFolds()함수 사용. #12-4