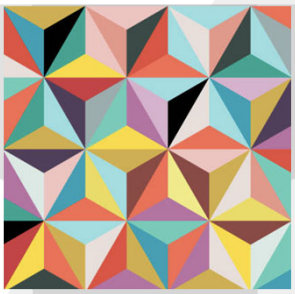
빅데이터 분석

[R 데이터 시각화 - ggplot 패키지]

데이터 시각화 기법 기초



데이터 시각화 이해

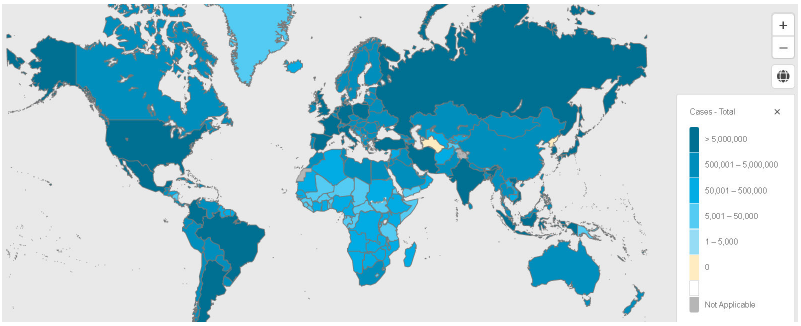
데이터 시각화란?

- 데이터 분석 결과를 쉽게 이해할 수 있도록 시각적으로 표현하는 과정
- 데이터를 요약하고, 한 눈에 살펴볼 수 있도록 돕는 시각화가 필수

시각화가 이루어지지 않은 데이터

#	Country, Other	Total Cases	New Cases	Total Deaths	New Deaths
	World	496,675,751	+389,236	6,195,655	+681
1	USA	81,988,278		1,011,096	
2	India	43,033,067		521,604	
3	Brazil	30,094,388		661,035	
4	France	26,549,263		143,017	
5	Germany	22,364,607		131,805	
6	UK	21,508,546		169,412	
7	Russia	17,955,120		370,889	
8	Italy	15,106,066		160,973	
9	S. Korea	14,983,694	+205,289	18,754	+373

시각화가 이루어진 데이터



데이터 시각화 이해

효과적인 데이터 시각화 필수 요소

정보 이해를 돕는 데이터 구조 파악
(자료특성 & 변수 개수 등)

적절한 데이터 표현 기법 사용
(막대, 원형, 산점도, 박스차트 등)

효과적인 시각화 디자인
(색상, 폰트, 디자인 등)

데이터 시각화 기법 기초

데이터 시각화 기법 기초 도구

- 빅데이터 분석을 통해 얻은 결과에서 의미 있는 정보를 요약하여 전하는 과정 필요
- 데이터 시각화를 이용해 새로운 인사이트 발견
- 실무에서 가장 많이 사용되는 시각화 도구
  - treemap(트리맵 생성)
  - symbols(버블차트 생성)
  - mosaicplot(모자이크 플롯 생성)
  - ggplot(데이터 시각화 통합 도구)

데이터 시각화 기법 기초

데이터 시각화 기법 기초 – 트리맵

- 트리맵(Tree map)은 사각 타일의 형태로 구성되며, 각 타일의 크기와 색으로 데이터에 담긴 정보를 표현
- 각각 타일 간에는 계층 구조가 있어, 데이터에 존재하는 계층 구조 표현 가능
- treemap 패키지를 설치하고, 패키지 속 GNI2014 데이터셋을 활용

** GNI 2014는 2014년도 세계 국가별 인구, 국민총소득(GNI), 소속 대륙 정보를 포함하는 데이터*

* Gross National Income

```
> head(GNI2014)
  iso3          country     continent population    GNI
3  BMU          Bermuda North America      67837 106140
4  NOR           Norway        Europe    4676305 103630
5  QAT            Qatar          Asia     833285  92200
6  CHE      Switzerland        Europe    7604467  88120
7  MAC Macao SAR, China          Asia     559846  76270
8  LUX       Luxembourg        Europe     491775  75990
>
```

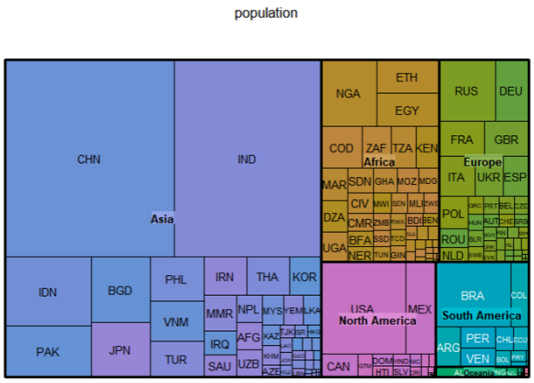
데이터 시각화 기법 기초

데이터 시각화 기법 기초 - 트리맵

treemap() 함수

treemap(data,
index=c("계층구조1", "계층구조2"),
vSize = "크기 기준변수")

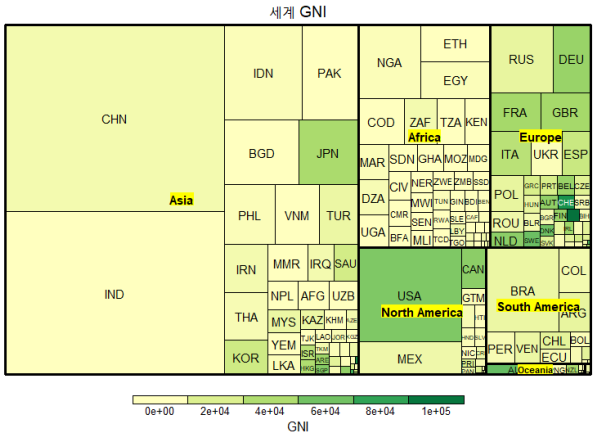
```
install.packages("treemap") #패키지 설치
library(treemap) #패키지 불러오기
data("GNI2014") #데이터 불러오기
head(GNI2014) #데이터 확인하기

treemap(GNI2014,
        index = c("continent", "iso3"), #계층구조 설정
        vSize = "population") #타일 크기
```



데이터 시각화 기법 기초

데이터 시각화 기법 기초 - 트리맵

```
install.packages("treemap") #패키지 설치
library(treemap) #패키지 불러오기
data("GNI2014") #데이터 불러오기
head(GNI2014) #데이터 확인하기

treemap(GNI2014,
        index = c("continent", "iso3"), #계층구조 설정
        vSize = "population", #타일 크기
        vColor = "GNI", #타일 색상
        type = "value", #타일 컬러링 방법
        bg.labels = "yellow", #레이블 배경색
        title = "세계 GNI") #트리맵 제목
```



데이터 시각화 기법 기초

데이터 시각화 기법 기초 - 트리맵

- GNI2014
  트리맵을 그릴 대상이 되는 데이터셋(데이터프레임 형태 입력)
- index = c("continent", "iso3")
  트리맵에서 타일을 대륙(continent) 안에 국가(iso3) 형태로 배치
- vSize = "population"
  타일의 크기를 결정하는 열을 지정 -> 본 코드에서는 인구 수(population) 기반 지정

* ISO 3166-1 Alpha-3은
ISO 3166-1에서 정한 알파벳 세 글자의 국가 코드로 나타낸 것

데이터 시각화 기법 기초

데이터 시각화 기법 기초 - 트리맵

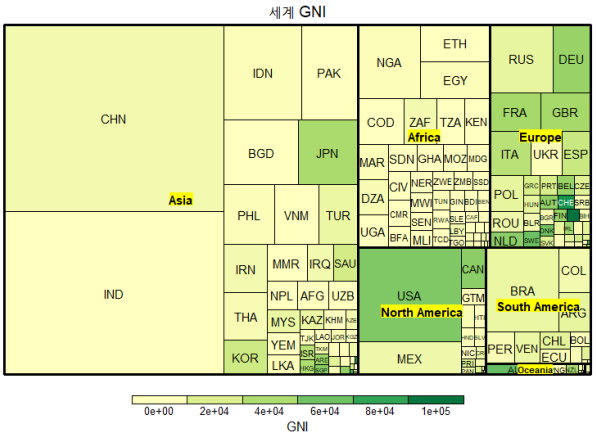
- vColor = "GNI"
  타일의 컬러링 방법을 지정 -> value는 vColor에서 지정한 열에 저장된 값의 크기에 따라 색상의 농도가 결정(index, comp, dens 등을 지정할 수 있음)
- bg.labels = "yellow"
  대륙을 나타내는 레이블의 배경색 지정
- title = "세계 GNI"
  트리맵의 제목을 지정

데이터 시각화 기법 기초

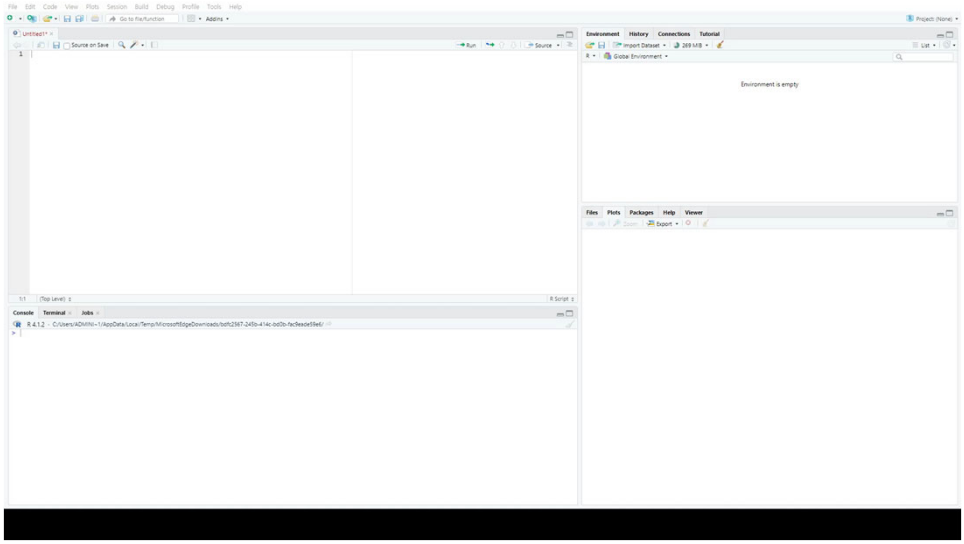
데이터 시각화 기법 기초 - 트리맵

```
install.packages("treemap") #패키지 설치
library(treemap) #패키지 불러오기
data("GNI2014") #데이터 불러오기
head(GNI2014) #데이터 확인하기

treemap(GNI2014,
        index = c("continent", "iso3"), #계층구조 설정
        vSize = "population", #타일 크기
        vColor = "GNI", #타일 색상
        type = "value", #타일 컬러링 방법
        bg.labels = "yellow", #레이블 배경색
        title = "세계 GNI") #트리맵 제목
```

인구 수를 바탕으로 타일 크기를 정하고,
GNI 수치에 따라 타일의 색을 칠한 트리맵 생성



데이터 시각화 기법 기초 - 트리맵

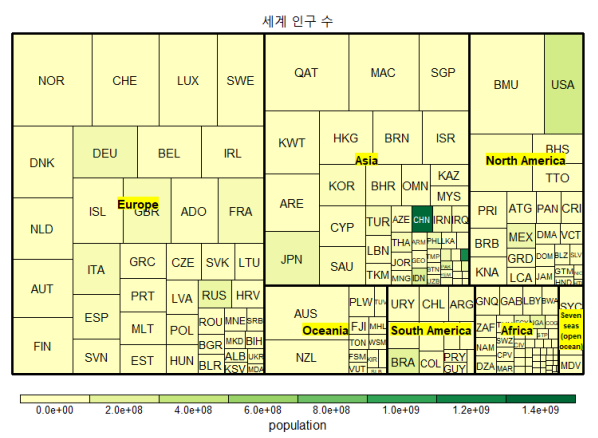


데이터 시각화 기법 기초

데이터 시각화 기법 기초 – 트리맵

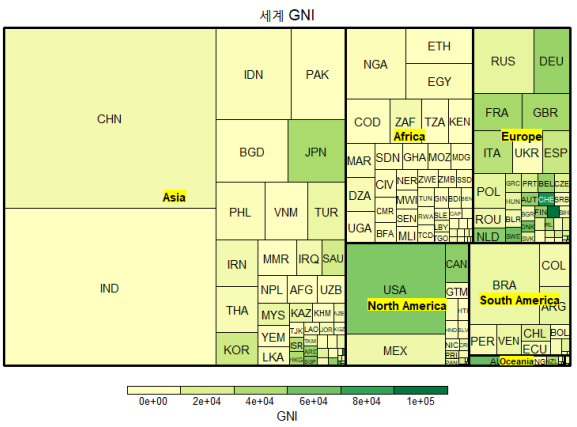
```
install.packages("treemap") #패키지 설치
library(treemap) #패키지 불러오기
data("GNI2014") #데이터 불러오기
head(GNI2014) #데이터 확인하기

treemap(GNI2014,
        index = c("continent", "iso3"), #계층구조 설정
        vSize = "GNI", #타일 크기
        vColor = "population", #타일 색상
        type = "value", #타일 컬러링 방법
        bg.labels = "yellow", #레이블 배경색
        title = "세계 GNI") #트리맵 제목
```

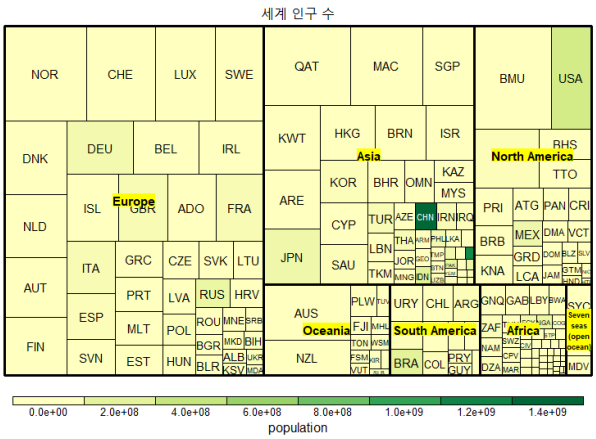
GNI 수치를 바탕으로 타일 크기를 정하고,
인구 수에 따라 타일의 색을 칠한 트리맵 생성



데이터 시각화 기법 기초

데이터 시각화 기법 기초 – 트리맵





데이터 시각화 기법 기초

실습 #1

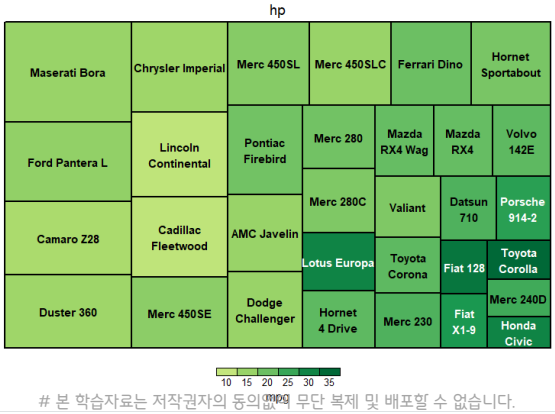
이제 취업을 앞둔 상명이는 중고차를 한 대 마련하려고 합니다. 그래서 먼저 중고차 시장에 있는 수많은 자동차 매물들을 한번에 시각적으로 확인할 수 있는 트리맵을 만들고자 합니다. 그 중에 엔진 출력과 연비를 중점으로 보고자 합니다.

treemap() 함수를 이용하여, 엔진 출력(hp)를 기준으로 크기(vSize)를 정하고 연비(mpg)를 기준으로 색상(vColor)을 칠하는 히트맵을 만드세요.

데이터 시각화 기법 기초

실습 #1

최종 결과물이 다음과 같이 출력되도록 하십시오.



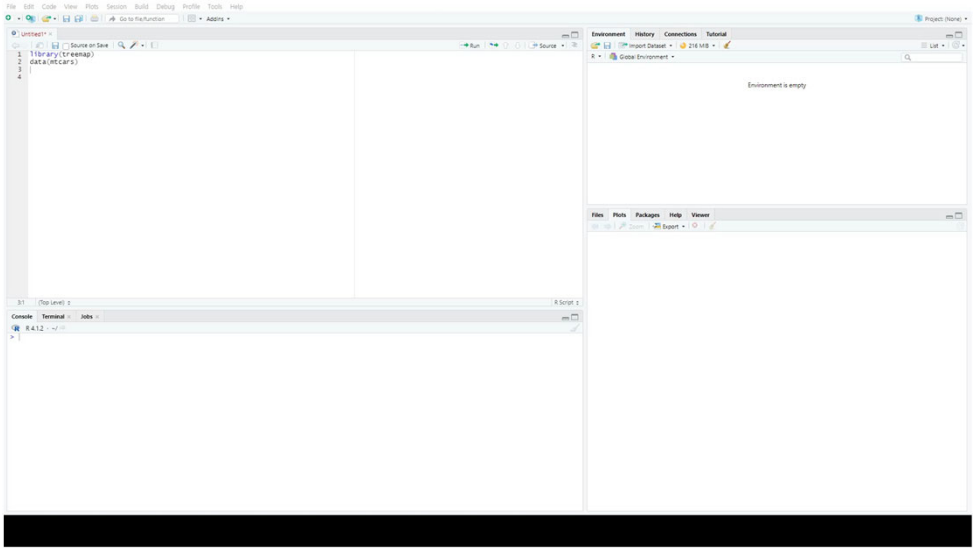
데이터 시각화 기법 기초

실습 #1

```
library(treemap)
data(mtcars)
mtcars$name <- rownames(mtcars)
treemap(mtcars,
        index = "name",
        vSize = "hp",
        vColor = "mpg",
        type = "value")
```

```
> library(treemap)
> data(mtcars)
> mtcars$name <- rownames(mtcars)
> treemap(mtcars,
+         index = "name",
+         vSize = "hp",
+         vColor = "mpg",
+         type = "value")
>
```


데이터 시각화 기법 기초 - 실습 #1



데이터 시각화 기법 기초

데이터 시각화 기법 기초 – 버블차트

- 버블차트(Bubble chart)는 산점도 위에 버블의 크기로 정보를 표시하는 시각화 방법
- 별도 패키지 설치 없이, R에서 바로 사용 가능
- state.x77 데이터셋을 사용

** state.x77는 미국의 주별 소득, 면적, 고교 졸업율, 살인율 등의 정보를 담은 데이터*

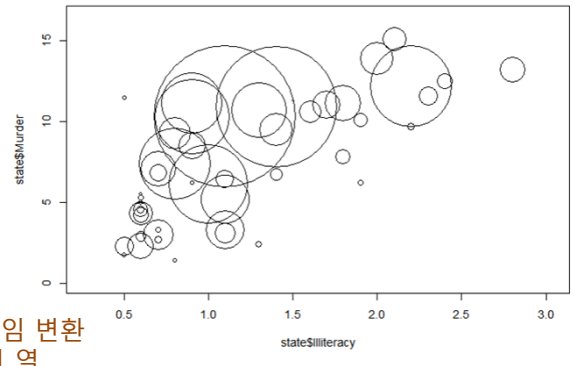
```
> head(state.x77)
           Population Income Illiteracy Life Exp Murder HS Grad Frost   Area
Alabama          3615   3624        2.1    69.05   15.1    41.3    20  50708
Alaska            365   6315        1.5    69.31   11.3    66.7   152 566432
Arizona          2212   4530        1.8    70.55    7.8    58.1    15 113417
Arkansas         2110   3378        1.9    70.66   10.1    39.9    65  51945
California      21198   5114        1.1    71.71   10.3    62.6    20 156361
Colorado         2541   4884        0.7    72.06    6.8    63.9   166 103766
> |
```


데이터 시각화 기법 기초

데이터 시각화 기법 기초 – 버블차트

symbols() 함수

symbols(x축 변수, y축 변수,
circles = 원 기준 변수)

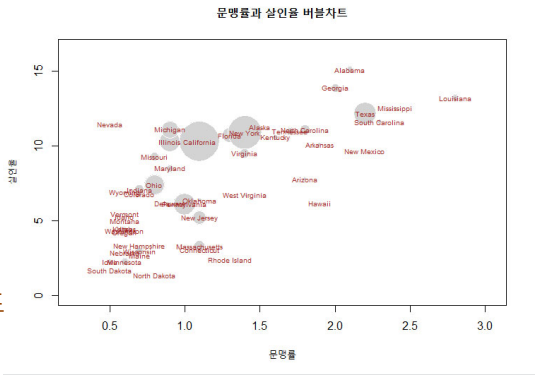



```
state <- data.frame(state.x77) #매트릭스 -> 데이터프레임 변환
symbols(state$Illiteracy, state$Murder, #원의 x, y 좌표의 열
        circles = state$Population) #원의 반지름 열
```

데이터 시각화 기법 기초

데이터 시각화 기법 기초 – 버블차트

```
state <- data.frame(state.x77) #매트릭스 -> 데이터프레임 변환
symbols(state$Illiteracy, state$Murder, #원의 x, y 좌표의 열
        circles = state$Population, #원의 반지름 열
        inches = 0.3, #원의 크기 조절 값
        fg = "white", #원의 테두리 색
        bg = "lightgray", #원의 바탕색
        lwd = 1.5, #원의 테두리 두께
        xlab = "문맹률", #x축 범례
        ylab = "살인율", #y축 범례
        main = "문맹률과 살인율 버블차트") #버블차트 제목

text(state$Illiteracy, state$Murder, #텍스트가 출력될 x, y 좌표
     rownames(state), #출력할 텍스트
     cex = 0.6, #폰트 크기
     col = "brown") #폰트 컬러
```



데이터 시각화 기법 기초

데이터 시각화 기법 기초 – 버블차트

symbols() 함수 argument

- state$Illiteracy, state$Murder
  2차원 좌표의 x축과 y축을 나타낼 열을 지정하여, x축 값과 y축 값이 만나는 지점에서 원을 생성
- circles = state$population
  원의 크기(반지름)을 결정할 열을 지정
- inches = 0.3
  원의 크기를 조절하는 매개변수로, 값이 클 수록 원이 크게 그려짐

데이터 시각화 기법 기초

데이터 시각화 기법 기초 - 버블차트

- fq = “white”
  원의 테두리 선 색을 지정
- bg = "lightgray
  원의 바탕색을 지정
- lwd = 1.5
  원의 테두리 선 두께를 지정
- xlab = “문맹률”
  x축의 범례를 지정

데이터 시각화 기법 기초

데이터 시각화 기법 기초 - 버블차트

- ylab = “살인율”
  y축의 범례를 지정
- main = “문맹률과 살인율 버블차트”
  버블차트의 제목을 지정

데이터 시각화 기법 기초

데이터 시각화 기법 기초 – 버블차트

text() 함수 argument

- state$Illiteracy, state$Murder
  텍스트를 표시할 위치에 대한 x, y축 좌표값을 나타내며, symbols() 함수에 있는 원의 x, y축 좌표값과 일치하도록 설정
- rownames(state)
  표시할 텍스트를 지정 -> 본 코드에서는 미국 각 주의 이름

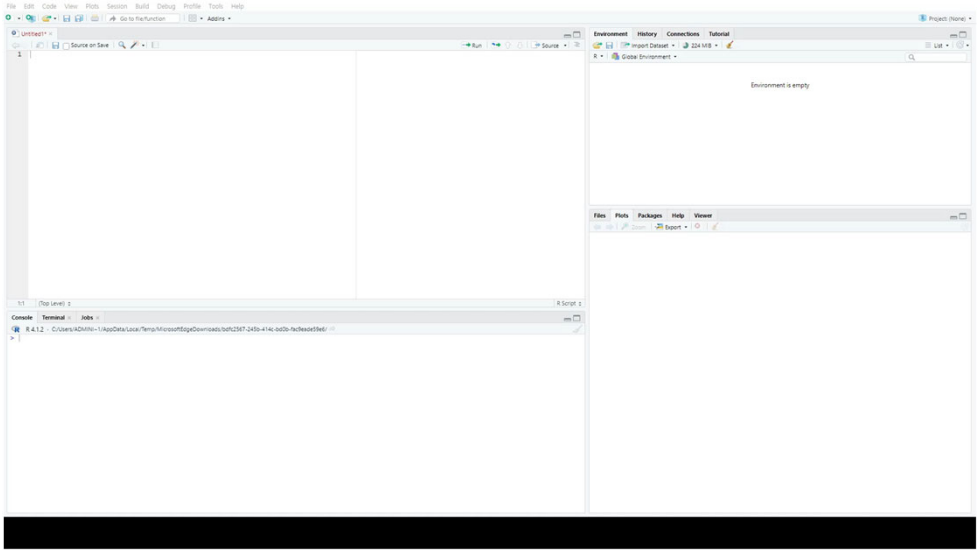
데이터 시각화 기법 기초

데이터 시각화 기법 기초 – 버블차트

text() 함수 argument

- cex = 0.6
  텍스트의 크기를 지정
- col = “brown”
  텍스트의 색을 지정

데이터 시각화 기법 기초 - 버블차트



데이터 시각화 기법 기초

실습 #2

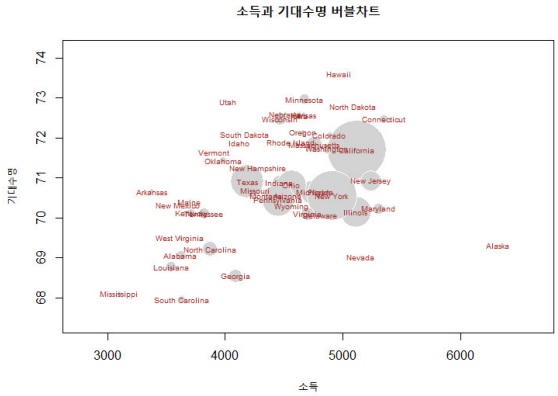
상명이가 자료조사를 하던 중, 미국 주별 소득과 기대수명 데이터를 찾아냈습니다.
이 데이터를 통해 미국 주별 소득과 기대수명 간의 관계를 버블차트를 통해 시각화를 해보려 합니다.

미국 주별 종합 데이터(state.x77)에서 소득(Income)과 기대수명(Life.exp) 데이터를 바탕으로, 1) 소득을 x축에 두고 2) 기대수명을 y축에 두며 3) 인구수로 버블 크기를 정하는 버블차트를 그려보세요.

데이터 시각화 기법 기초

실습 #2

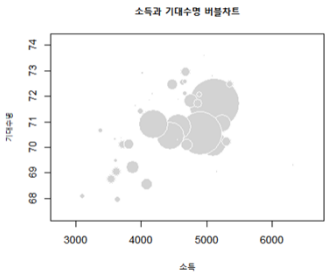
최종 결과물이 다음과 같이 출력되도록 하십시오.



데이터 시각화 기법 기초

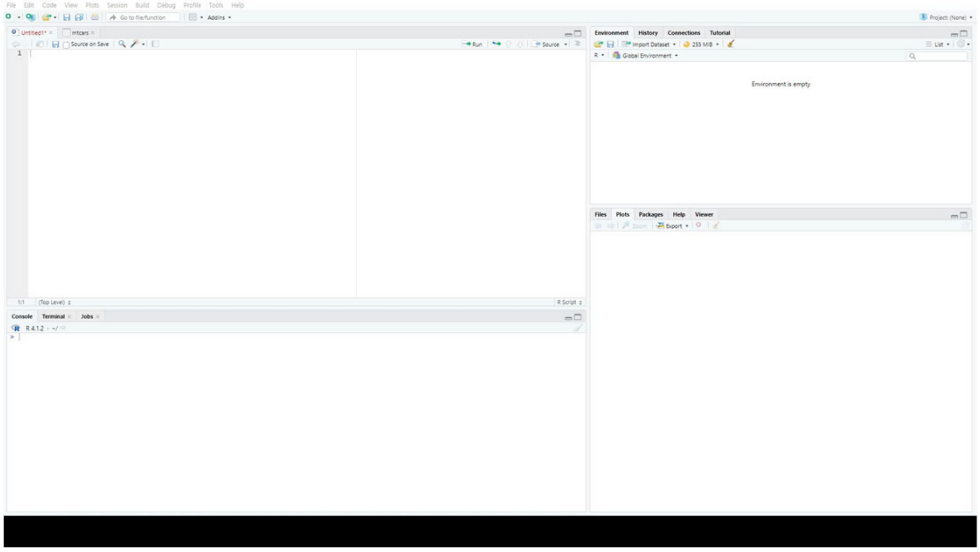
실습 #2

```
state <- data.frame(state.x77)
symbols(state$Income, state$Life.Exp,
        circles = state$Population,
        inches = 0.4,
        fg = "white",
        bg = "lightgray",
        lwd = 1.5,
        xlab = "소득",
        ylab = "기대수명",
        main = "소득과 기대수명 버블차트")
```




```
text(state$Income, state$Life.Exp,
     rownames(state),
     cex = 0.6,
     col = "brown")
```

데이터 시각화 기법 기초 - 실습 #2



데이터 시각화 기법 기초

데이터 시각화 기법 기초 - 모자이크 플롯

- 모자이크 플롯(Mosaic plot)은 다중변수 범주형 데이터에 대해 각 변수의 그룹별 비율을 면적으로 표시하여 정보를 전달
- R에서 제공하는 mtcars 데이터셋 사용

  ** mtcars는 32개의 차종들의 각각 차량 성능(기어, 연비, 중량, 마력, 기통 등)을 담은 데이터*

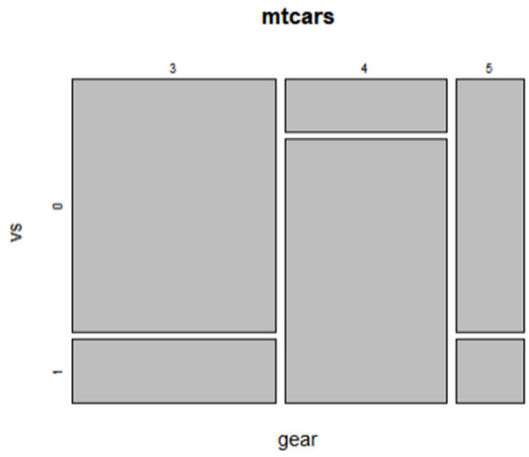
데이터 시각화 기법 기초

데이터 시각화 기법 기초 – 모자이크 플롯

mosaicplot() 함수

mosaicplot(~ x축 변수 + y축 변수,
data = 데이터프레임)

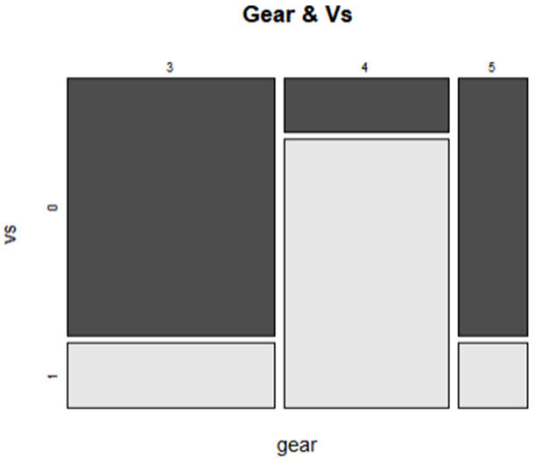
```
data(mtcars)

mosaicplot(~gear+vs,
           data = mtcars)
```



데이터 시각화 기법 기초

데이터 시각화 기법 기초 – 모자이크 플롯

```
data(mtcars)

mosaicplot(~gear+vs,
           data = mtcars,
           color = TRUE,
           main = "Gear & Vs")
```



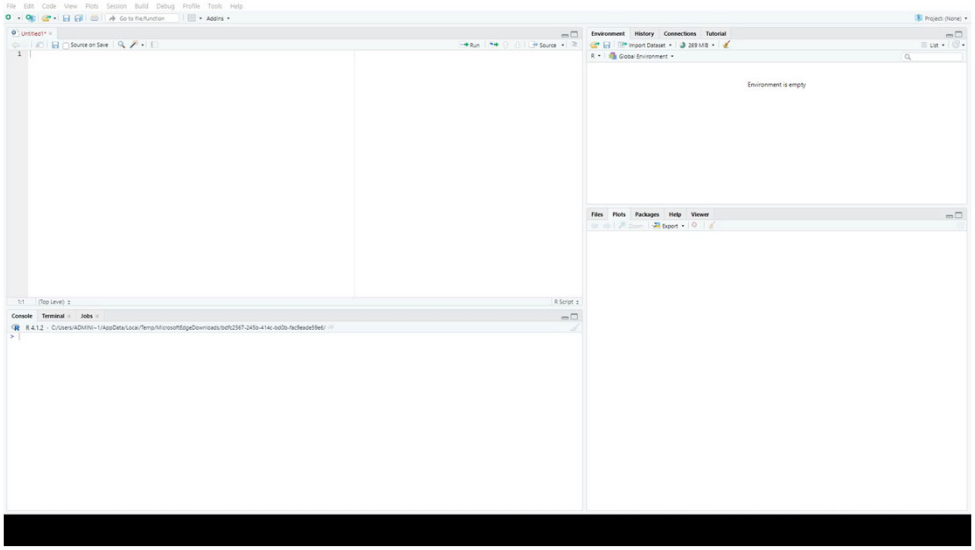
데이터 시각화 기법 기초

데이터 시각화 기법 기초 – 모자이크 플롯

mosaicplot() 함수 argument

- ~gear+vs
  모자이크 플롯을 작성할 대상 변수 지정(gear = 변속단수, vs = 신기술 엔진 유무)
  ~다음 변수 = x축 방향으로 표시, + 다음 변수 = y축 방향으로 표시
- data = mtcars
  모자이크 플롯을 작성할 대상 데이터셋 지정
- color = TRUE
  y축 변수를 그룹별로 음영을 다르게 표시
- main = “Gear & Vs”
  모자이크 플롯의 제목을 지정

데이터 시각화 기법 기초 – 모자이크 플롯



데이터 시각화 기법 기초

실습 #3

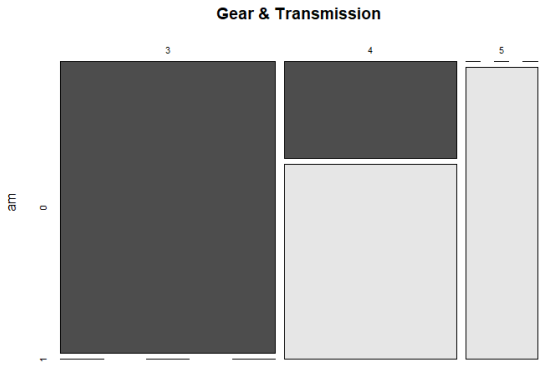
상명이는 중고차 시장으로 다시 돌아와, 이번에는 중고차들 중 기어 변속단수(gear)과 자동변속 여부(am)의 분포를 확인하고자 합니다. 이를 위해 간단한 모자이크 플롯으로 시각화를 하려 합니다.

mtcars 데이터셋을 활용하여 기어 변속단수를 x축으로 하고, 자동변속 여부를 y축으로 하는 모자이크 플롯을 그리시오.

데이터 시각화 기법 기초

실습 #3

모자이크 플롯의 결과물은 다음과 같아야 합니다.

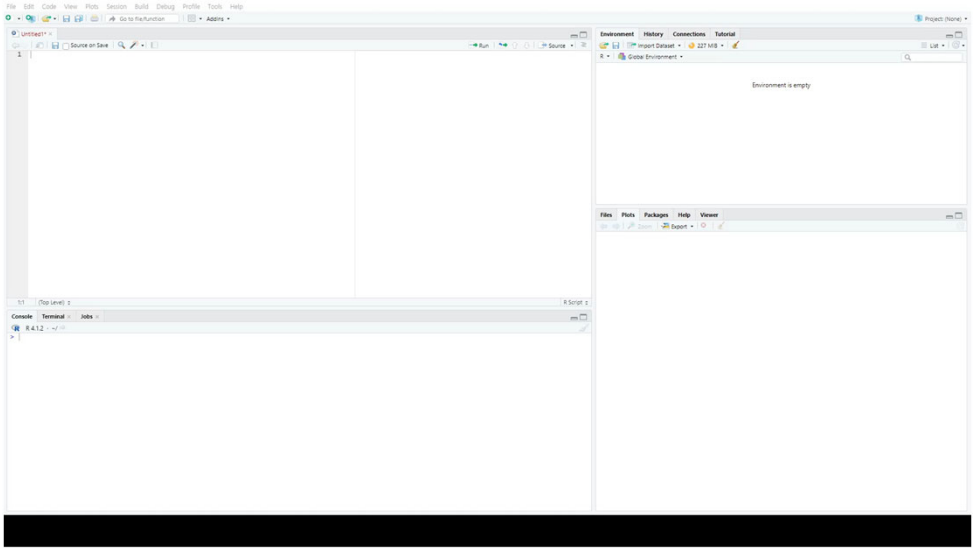


데이터 시각화 기법 기초

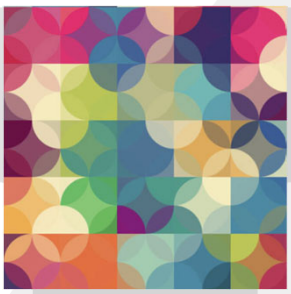
실습 #3

```
data(mtcars)
mosaicplot(~gear+am,
        data = mtcars,
        color = TRUE,
        main = "Gear & Transmission")
```


데이터 시각화 기법 기초 - 실습 #3



ggplot 시각화



ggplot 시각화

ggplot 개념 ** Grammar of Graphics - Plot*

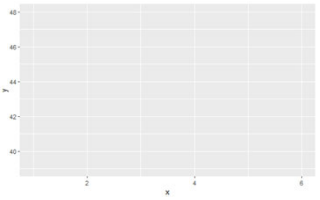
- 데이터 분석에서 기본 함수를 사용해 그래프를 그릴 수 있지만, 더욱 심미적인 그래프 작업을 위해 ggplot 패키지를 주로 사용
- ggplot으로 데이터 시각화를 세부적으로 할 수 있으나, 복잡해질 수 있다는 단점 존재

```
install.packages("ggplot2")
library(ggplot2)
```

* ggplot2패키지는 ggplot패키지의 업데이트된 버전

```
> install.packages("ggplot2")
WARNING: Rtools is required to build R packages but is not currently installed. Please
 download and install the appropriate version of Rtools before proceeding:

https://cran.rstudio.com/bin/windows/Rtools/
'C:/Users/User/Documents/R/win-library/4.1'의 위치에 패키지(들)을 설치합니다.
(왜냐하면 'lib'가 지정되지 않았기 때문입니다)
trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.1/ggplot2_3.3.6.zip'
Content type 'application/zip' length 4128311 bytes (3.9 MB)
downloaded 3.9 MB

package 'ggplot2' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\User\AppData\Local\Temp\RtmpohNoN9\downloaded_packages
> library(ggplot2)
>
```

ggplot 시각화



ggplot 기본구조

* aes : aesthetic

ggplot(data = 데이터프레임, aes(x=x축 변수, y=y축 변수)) # 그래프 틀

+ <geom_FUNCTION>(stat="identity", width =수치, fill = " 컬러") # 그래프 형태

aes = 그래프의 미적 부분으로 x축, y축, 컬러 등을 지정하는 작업
- x: X축
- y: Y축
- alpha: 투명도
+ color: 그래프의 색깔, 모양일 경우 테두리
+ fill: 채우는 색깔
+ size: 선 굵기 또는 점의 크기
+ linetype: 선 패턴

geom_FUNCTION
ggplot으로부터 데이터를 받아, 실제 그래프를 그리는 작업
- geom_bar: 막대 그래프
- geom_line: 선 그래프
- gom_point: 산점도
- geom_boxplot : 상자 그래프

ggplot 시각화

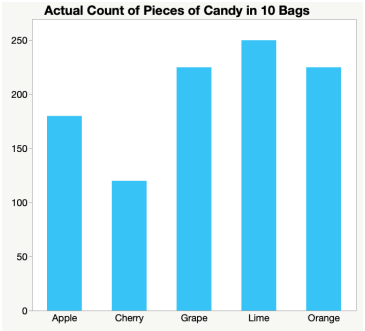
ggplot 주요 그래프 종류

막대 그래프	선 그래프	산점도	상자 그래프

ggplot 시각화

ggplot 시각화 – 막대 그래프

- 범주형 변수에 대한 값의 빈도 개수를 표시하는 그래프
- 가장 많이 사용되는 기본 그래프로, 각 변수 별로 값을 비교하는데 사용

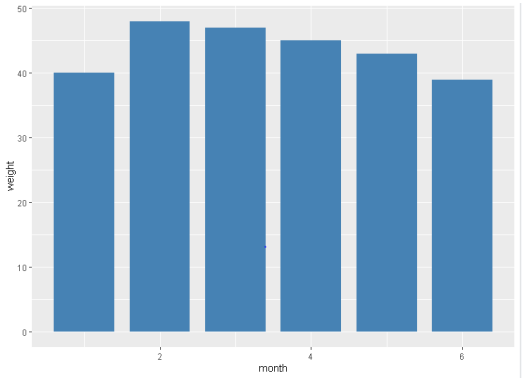


ggplot 시각화

ggplot 시각화 – 막대 그래프

월별 몸무게 데이터로 간단한 막대 그래프 생성

```
library(ggplot2) #ggplot 패키기 가져오기
month <- c(1, 2, 3, 4, 5, 6) #월 벡터
weight <- c(40, 48, 47, 45, 43, 39) #몸무게 벡터
df <- data.frame(month, weight) #월 & 몸무게 데이터프레임

ggplot(data=df, aes(x = month, y = weight)) #x, y축 지정
+ geom_bar(stat = "identity", width = 0.8, fill = "steelblue")
#막대 높이, 막대 폭, 막대 색상 지정하여 그래프 생성
```



ggplot 시각화

ggplot 시각화 - 막대 그래프

ggplot() 함수 argument

- df
  그래프를 작성할 데이터가 저장된 데이터프레임 지정
  행렬의 경우, data.frame() 사용해서 데이터프레임으로 변환하여 입력
- aes(x = month, y = weight)
  aes로 그래프를 그릴 때 사용할 x, y축의 열을 지정

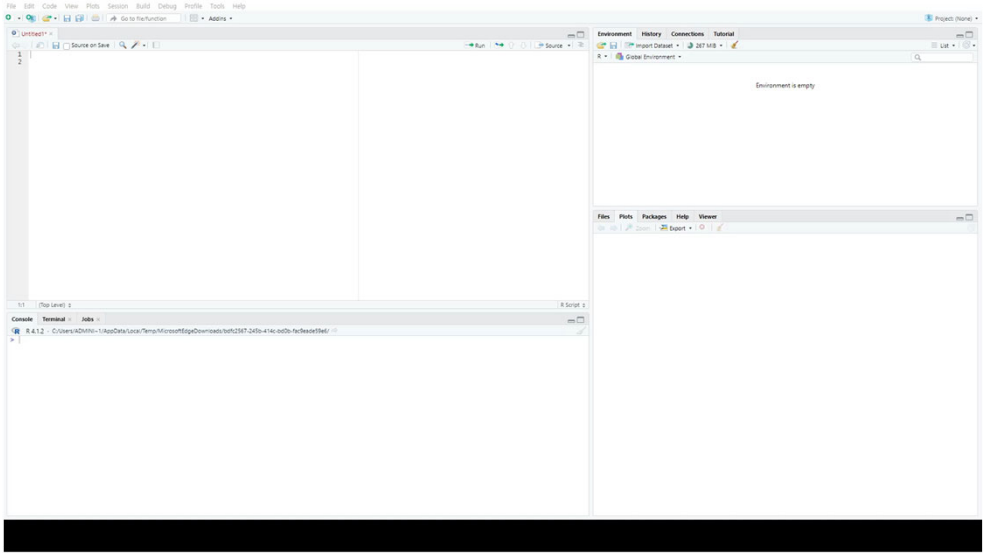
ggplot 시각화

ggplot 시각화 - 막대 그래프

geom_bar() 함수 argument

- stat = “identity”
  막대의 높이는 ggplot() 함수에서 y축에 해당하는 열에 의해 결정되도록 지정
  * stat = “count”이면 y축의 높이를 데이터 빈도(count)로 표시
- width = 0.8
  막대의 폭을 지정
- fill = “steelblue”
  막대 색상을 지정

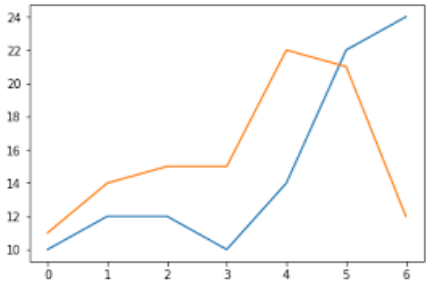
ggplot 시각화 - 막대 그래프



ggplot 시각화

ggplot 시각화 - 선 그래프

- 점들을 선분으로 이어 그린 그래프
- 시간에 따라 무언가가 지속적으로 변화하는 것을 기록할 때 유용
- 조사하지 않은 중간의 값도 대략 예측할 수 있다는 장점



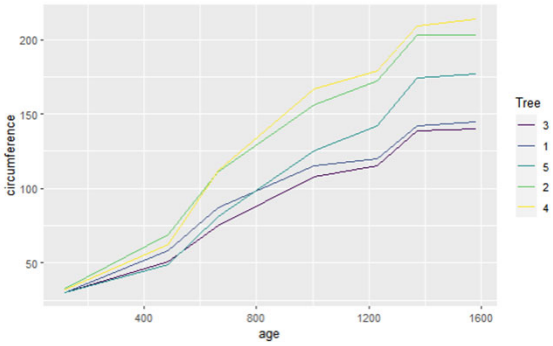
ggplot 시각화

ggplot 시각화 - 선 그래프

Orange 데이터셋의 나무 성장 그래프
(1부터 5까지의 나무의 나이에 따른 성장을 나타낸 데이터)

```
library(ggplot2)
head(Orange)

ggplot(data= Orange,
    aes(x = age, y = circumference)) # 그래프 작성 대상 지정
+ geom_line(aes(color = Tree)) # 선그래프 생성
```

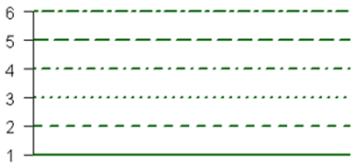
```
  Tree  age circumference
1    1  118            30
2    1  484            58
3    1  664            87
4    1 1004           115
5    1 1231           120
6    1 1372           142
```

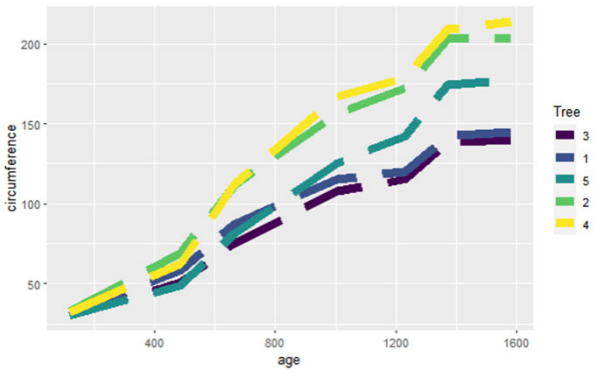


ggplot 시각화

ggplot 시각화 - 선 그래프

```
library(ggplot2)
head(Orange)

ggplot(data= Orange,
    aes(x = age, y = circumference))
+ geom_line(aes(color = Tree), linetype = 5, size=3)
```





ggplot 시각화

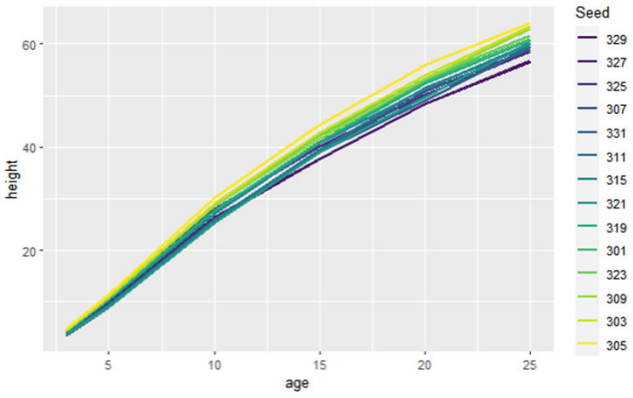
실습 #4

로볼리 소나무 생육 데이터를 모아둔 데이터셋 (Loblolly)이 있습니다.
만일 로볼리 소나무의 성장에 따른 크기를 시각화 하려할 때, 종자별 R 코드를 이용하여 시각화 합니다.

ggplot 시각화

실습 #4

로볼리 데이터셋 시각화 결과물은 다음과 같아야 합니다.

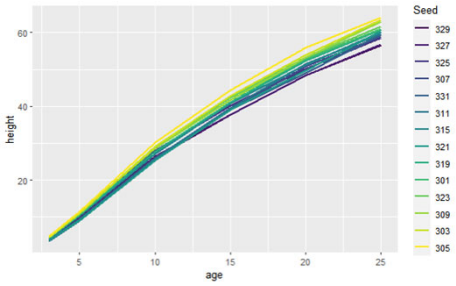


ggplot 시각화

실습 #4

```
data(Loblolly)
head(Loblolly)
install.packages("ggplot2")
library(ggplot2)

ggplot(data= Loblolly, aes(x = age, y = height))
+ geom_line(aes(color = Seed), linetype= 1, size=1)
```

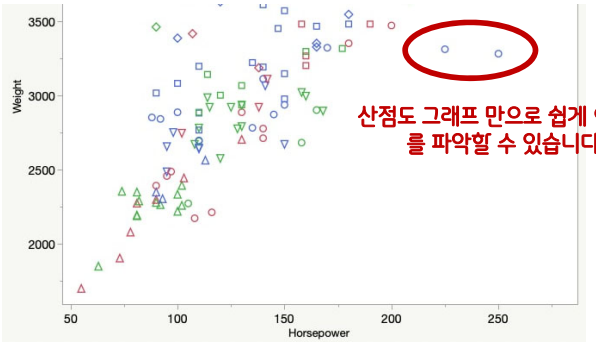
```
> data(Loblolly)
> head(Loblolly)
   height age Seed
1    4.51   3  301
15  10.89   5  301
29  28.72  10  301
43  41.74  15  301
57  52.70  20  301
71  60.92  25  301
>
```



ggplot 시각화

ggplot 시각화 – 산점도

- 2개 이상의 연속형 변수를 가질 때, 변수 간의 상관관계를 표현하기 위해 사용
- 상관관계 파악 뿐만 아니라, 평균에서 벗어난 이상치 값 파악을 할 때 유용

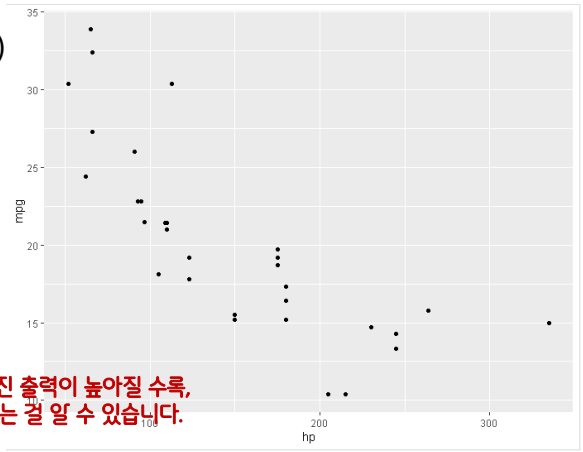


ggplot 시각화

ggplot 시각화 - 산점도

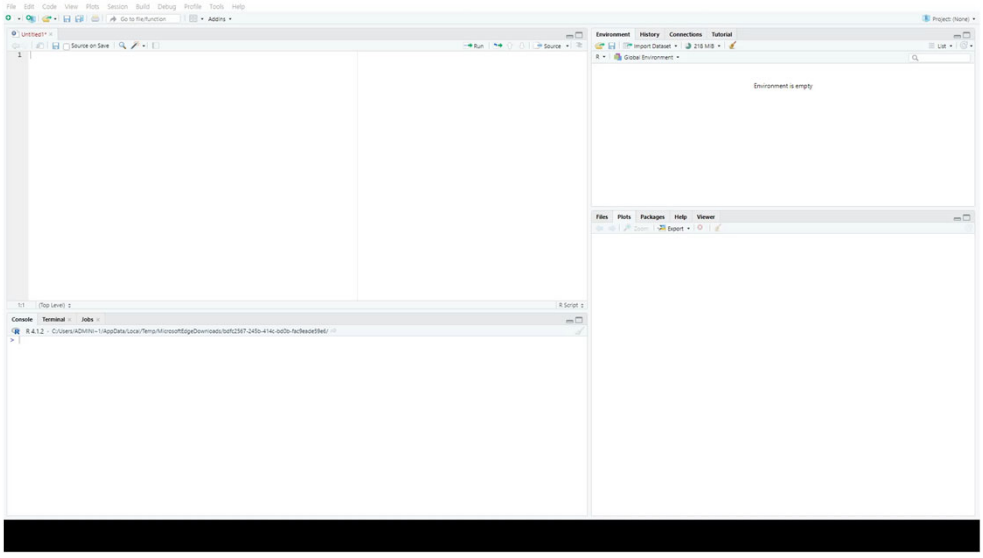
mtcars 데이터셋의 hp(엔진 출력)와 mpg(연비)의 상관관계를 보여주는 산점도 생성

```
data(mtcars)
library(ggplot2)

ggplot(data=mtcars, aes(x = hp, y = mpg))
+ geom_point()
```



산점도를 통해 엔진 출력이 높아질 수록,
연비가 낮아진다는 걸 알 수 있습니다.

ggplot 시각화 - 산점도



ggplot 시각화

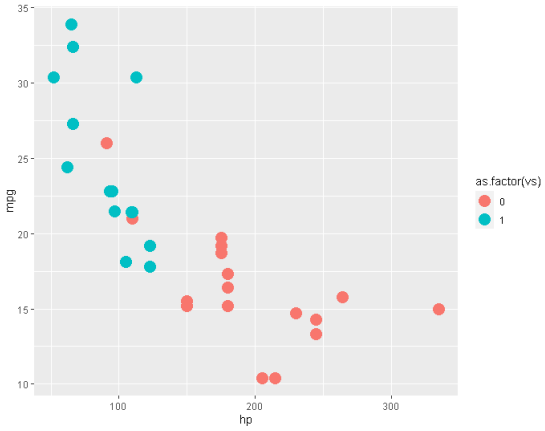
ggplot 시각화 - 산점도

```
> head(mtcars)
                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
> |
```

vs(신기술 엔진 여부)를 바탕으로, hp(엔진 출력)와 mpg(연비)의 상관관계를 보여주는 산점도 생성

```
data(mtcars)
library(ggplot2)

ggplot(data=mtcars, aes(x = hp, y = mpg))
+ geom_point(aes(color = as.factor(vs)))
```

* as.factor : data의 type을 factor로 변환하는 함수



ggplot 시각화

실습 #5

의료용 특수 절삭 기구를 제작하기 위해서 다이아몬드를 사러 의료기기점에 갔습니다. 그러나 수많은 다이아몬드들을 보고서 다이아몬드 캐럿(크기), 가격, 품질 간의 관계에 대해서 궁금해지기 시작했습니다.

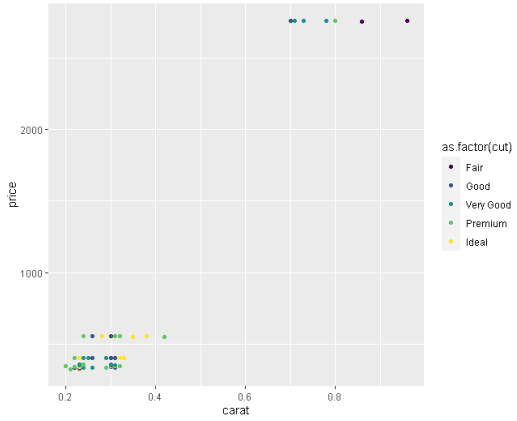
diamonds 데이터셋을 이용해, 다이아몬드 캐럿(carat)을 x축으로 하고, 가격(price)을 y축으로 하며, 품질(cut)에 따라 색상을 다르게 표기하는 산점도를 그리시오.

** 데이터가 매우 큰 관계로 diamonds <- head(diamonds, 100)를 통해 데이터의 개수를 100개로 제한하여 사용하시길 바랍니다.*

ggplot 시각화

실습 #5

산점도 결과물은 다음과 같이 나타나야 합니다.

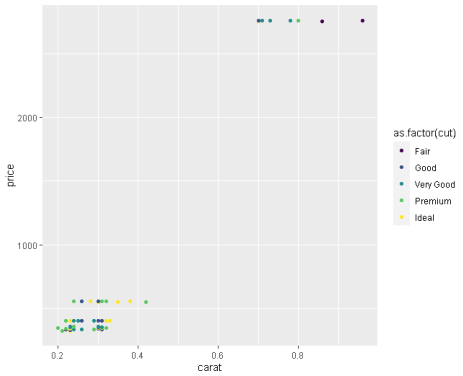


ggplot 시각화

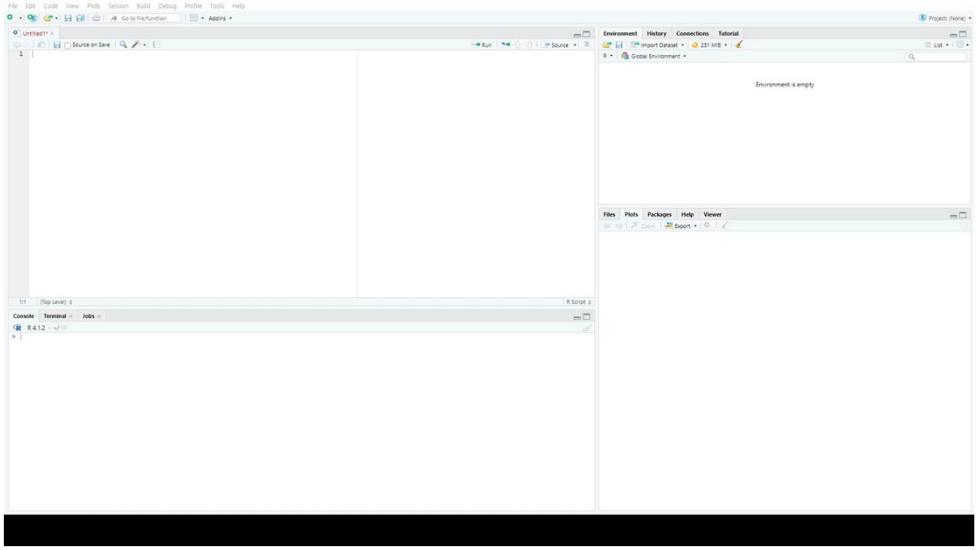
실습 #5

```
library(ggplot2)
data(diamonds)

diamonds <- head(diamonds, 100)

ggplot(data=diamonds, aes(x = carat, y = price))
+ geom_point(aes(color = as.factor(cut)))
```

```
> head(diamonds)
# A tibble: 6 x 10
  carat cut       color clarity depth table price     x     y     z
  <dbl> <ord>     <ord> <ord>   <dbl> <dbl> <int> <dbl> <dbl> <dbl>
1  0.23 Ideal     E     SI2      61.5    55   326  3.95  3.98  2.43
2  0.21 Premium   E     SI1      59.8    61   326  3.89  3.84  2.31
3  0.23 Good      E     VS1      56.9    65   327  4.05  4.07  2.31
4  0.29 Premium   I     VS2      62.4    58   334  4.2   4.23  2.63
5  0.31 Good      J     SI2      63.3    58   335  4.34  4.35  2.75
6  0.24 Very Good J     VVS2     62.8    57   336  3.94  3.96  2.48
>
```



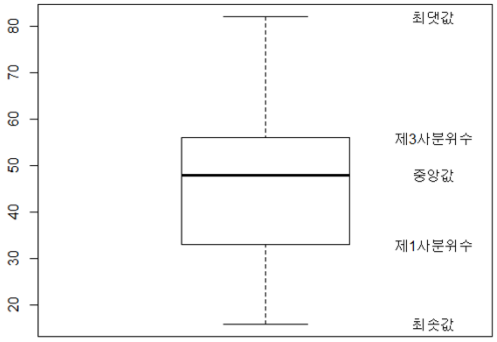
ggplot 시각화 - 실습 #5



ggplot 시각화

ggplot 시각화 - 상자 그래프

- 연속형 변수에 대한 데이터 분포를 표시하는 그래프
- 백분위수(25, 50, 75번째)와 최소 & 최대값을 표시하며, 분포와 이상치를 찾는데 특화

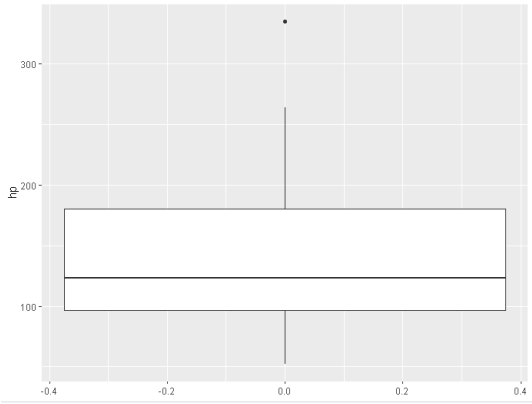


ggplot 시각화

ggplot 시각화 - 상자 그래프

mtcars 데이터셋의 hp(엔진 출력)의 백분위수 및 최대 & 최소값을 보여주는 상자 그래프 생성

```
library(ggplot2)

ggplot(data=mtcars, aes(y = hp)) + geom_boxplot()
```

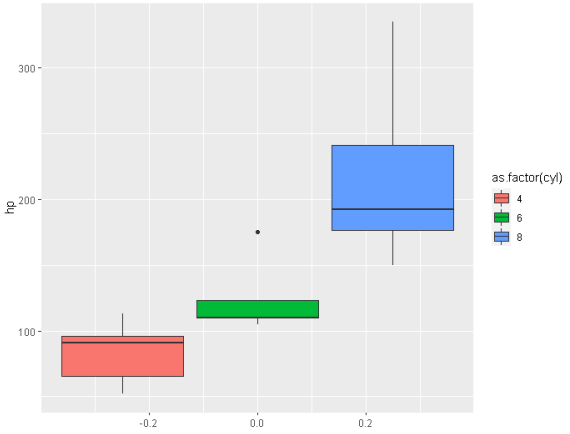


ggplot 시각화

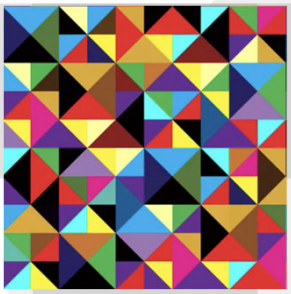
ggplot 시각화 - 상자 그래프

cyl(엔진 기통 개수)를 바탕으로, hp(엔진 출력)의 분포를 보여주는 상자 그래프 생성

```
library(ggplot2)

ggplot(data=mtcars, aes(y = hp))
+ geom_boxplot(aes(fill = as.factor(cyl)))
```



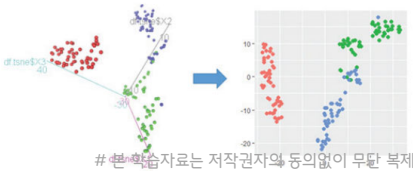
차원 축소



차원축소

개념

- 산점도는 2차원 평면 상에 두 변수의 값으로 좌표를 정하여 위치를 나타내는 방법
- 변수가 4개인 4차원 데이터에 대한 산점도를 어떻게 그릴 수 있는가?
  - 현재로서 4차원을 시각화 할 수는 없음
  - 4차원을 2차원으로 축소하여 그리는 방법으로 해결
- 차원축소(Dimension reduction)란 고차원 데이터를 2~3차원으로 축소하는 기법
  - 차원축소 이유 : 2~3차원으로 축소된 데이터로 산점도를 작성하여 데이터 분포 확인을 위해
  - 차원축소 방법 : 3차원 상의 물체를 빛에 비추면 그림자가 생기는 것과 비슷한 방법 활용



3차원 상의 데이터 분포를 2차원 상의 분포로 변환
마치 투명한 육면체 위에 표시된 점들을 사진으로 찍은 것처럼 구현

차원축소

4차원 데이터를 2차원으로 축소

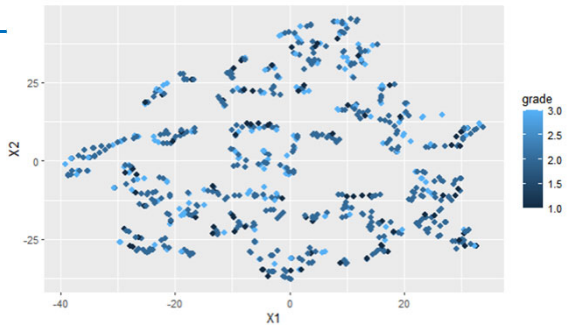
- 각종 질병 환자 데이터를 담은 surivival 패키지를 불러와서, 유방암 환자 관련 데이터셋 gbsg 데이터셋을 활용
  - gbsg(German Breast Cancer Study Group) 패키지는 실제 독일에서 1984-1989년 686명의 유방암 환자들의 ID, 연령, 폐경기 상태, 종양 크기, 전이 단계, 림프절 수, 호르몬 요법 적용 여부, 생존기간, 생존여부 등을 수집하여 정리한 데이터
  - 이중 연령(age), 종양 크기(size), 전이 단계(grade), 림프절 수(node) 데이터를 사용
- 차원 축소를 통해, 4차원 -> 2차원으로 변환하는 과정 실습

차원축소

Rtsne() 함수

- t-SNE(t-distribution Stochastic Neighbor Embedding)는 N차원에 분포된 데이터들의 거리 정보를 보존하면서, 차원을 축소하는 방법
- Rtsne 패키지의 Rtsne() 함수는 N차원 데이터프레임 데이터를 차원축소
- Rtsne(데이터프레임, dim = 차원 수, perplexity = 10, check_duplicate = F)
  - 만일 mtcars 데이터셋을 4차원에서 2차원으로 차원축소 할 경우
    tsne(mtcars, dim = 2, perplexity = 10, check_duplicate = F)

차원축소



4차원을 2차원으로 축소

```
library(Rtsne)
library(ggplot2)
library(survival)
cancer <- gbsg[, c("age", "size", "grade", "nodes")]  #gbsg 데이터셋 중 필요한 열만 추출
grade <- cancer$grade  #grade 열 추출

#4차원 데이터를 2차원으로 축소하기
tsne <- Rtsne(cancer, dim = 2, perplexity = 10, check_duplicates = FALSE)

#차원축소 결과 시각화
tsneDF <- data.frame(tsne$Y)
ggplot(data=tsneDF, aes(x=X1, y=X2, color = grade)) + geom_point(size = 2)
```


차원축소

1. 라이브러리 불러오기

```
library(Rtsne)
library(ggplot2)
library(survival)
```

2. 유방암 환자 관련 gbsg 데이터셋 가져와서 필요한 열만 추출하기

```
cancer <- gbsg[, c("age", "size", "grade", "nodes")]  #gbsg 데이터셋 중 필요한 열만 추출
grade <- cancer$grade  #grade 열 추출
```

차원축소

3. Rtsne 함수를 활용하여 차원 축소하기

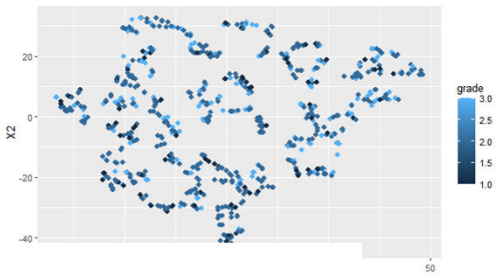
tsne <- Rtsne(cancer, dim = 2, perplexity = 10, check_duplicates = FALSE)

4. 차원축소 결과를 시각화하기

tsneDF <- data.frame(tsne$Y)

ggplot(data=tsneDF, aes(x=X1, y=X2, color = grade)) + geom_point(size = 2)

차원축소




```
> library(Rtsne)
> library(ggplot2)
> library(survival)
> cancer <- gbsg[, c("age", "size", "grade", "nodes")]  #gbsg 데이터셋 중 필요
한 열만 추출
> grade <- cancer$grade  #grade 열 추출
>
> #4차원 데이터를 2차원으로 축소하기
> tsne <- Rtsne(cancer, dim = 2, perplexity = 10, check_duplicates = FALSE)
>
> #차원축소 결과 시각화
> tsneDF <- data.frame(tsne$Y)
> ggplot(data=tsneDF, aes(x=X1, y=X2, color = grade)) + geom_point(size = 2)
>
```

SUMMARY

- 데이터 시각화
  - 트리맵 treemap() 함수
    - ✓ treemap(data, index=c("계층구조1", "계층구조2"), vSize="크기 기준변수")
  - 버블차트 symbols() 함수
    - ✓ symbols(x축 변수, y축 변수, circles = 원 기준 변수)
  - 모자이크 플롯 mosaicplot() 함수
    - ✓ mosaicplot(~ x축 변수 + y축 변수, data = 데이터프레임)

SUMMARY

- ggplot

  ggplot(data = 데이터프레임, aes(x=x축 변수, y=y축 변수)) # 그래프 틀

  + <geom_FUNCTION>(stat="identity", width =수치, fill = " 컬러") # 그래프 형태

  - geom_FUNCTION
    - ✓ geom_bar: 막대 그래프
    - ✓ geom_line: 선 그래프
    - ✓ gom_point: 산점도
    - ✓ geom_boxplot: 상자 그래프

SUMMARY

- **차원 축소**
  - Rtsne(데이터프레임, dim = 차원 수, perplexity = 10, check_duplicate= F)

감사합니다.