tidyverse

한상곤(sangkon@pusan.ac.kr)

2023.06.13(화)

Contents

1 tidyverse

`tidyverse`는 R에서 데이터분석에 필수적으로 사용되는 패키지입니다. RStudio의 해들리 위캠(Hadley Wickham) 박사 팀이 주도적으로 이끌고 있으며 대표적으로 `ggplot2`, `dplyr`, `tidyr`, `readr`, `purrr`, `tibble`, `stringr`, `forcats` 등이 포함되어 있습니다. 단일 패키지가 아니라 통합 패키지 형태로 되어 있이서, 한번에 설치가 가능합니다.

- ggplot은 데이터를 시각화하는데 사용
- dplyr는 `data.frame` 기반의 데이터 처리에 사용
- tidyr는 조건에 따른 데이터 처리에 사용
- readr는 tableau 데이터를 읽는데 사용
- purrr는 데이터를 일괄처리하는데 사용
- tibble은 `data.frame`과 유사한 자료구조를 제공
- stringr은 문자열을 다루기 위해서 사용
- forcats은 범주형 자료를 다룰 때 사용

1.1 tidyverse 설치

`install.packages`를 사용해서 설치하면 관련된 모든 패키지가 설치됩니다.

```
if(!require(tidyverse)) install.packages("tidyverse")
```

```
── Attaching core tidyverse packages ──────────────────── tidyverse 2.0.0 ──
 dplyr     1.1.2      readr     2.1.4
 forcats   1.0.0      stringr   1.5.0
 ggplot2   3.4.2      tibble    3.2.1
 lubridate 1.9.2      tidyr     1.3.0
 purrr     1.0.1
── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()
 Use the conflicted package to force all conflicts to become errors
```

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
```

```
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.2     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.1
## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

1.2 dplyr

데이터를 정리하기 위해서 dplyr에서 제공하는 변환은 크게 5가지입니다.

- 행
  - 선택
    - filter, 데이터에서 특정 열의 값이 조건에 맞는 행을 선택
    - slice, 데이터에서 특정 위치의 행을 선택
  - 정렬
    - arrange, 특정 열의 값을 기준으로 데이터의 행을 정렬
- 열
  - 선택
    - select, 열의 이름, 위치, 데이터 형식 등으로 열을 선택
  - 추가
    - mutate, 기존 열을 사용하여 새로운 열을 데이터에 추가
- 요약
  - summarize(), 데이터 전체 또는 특정 열을 하나의 통계량으로 요약
  - group_by(): 데이터 요약에만 사용되는 것은 아니나, 그룹별로 데이터를 통계 요약할 때 자주 사용

mpg 데이터를 사용해서 예제를 진행하겠습니다. mpg는 1999년과 2008년에 미국 EPA에서 조사하여 발표한 자동차 주요 모델별 연비 데이터입니다. mpg 데이터는 234 개의 행이 있으며, 각 행은 다음과 같은 변수로 구성되어 있습니다.

- manufacturer: 자동차 제조사
- model: 자동차 모델명
- displ: 자동차 배기량
- year: 제조년도
- cyl: 엔진 실린더 수
- trans: 자동차 트랜스미션 종류
- drv: 자동차 구동 방식. f=전륜구동, r=후륜구동, 4=사륜구동
- cty: 도심 연비 (마일/갤론)
- hwy: 고속도로 연비 (마일/갤론)
- fl: 연료 종류
- class: 자동차 분류

```
data(mpg)
mpg
```

```
## # A tibble: 234 x 11
##    manufacturer model      displ  year   cyl trans drv     cty   hwy fl    class
##    <chr>        <chr>      <dbl> <int> <int> <chr> <chr> <int> <int> <chr> <chr>
##  1 audi         a4           1.8  1999     4 auto~ f        18    29 p     comp~
##  2 audi         a4           1.8  1999     4 manu~ f        21    29 p     comp~
##  3 audi         a4           2    2008     4 manu~ f        20    31 p     comp~
##  4 audi         a4           2    2008     4 auto~ f        21    30 p     comp~
##  5 audi         a4           2.8  1999     6 auto~ f        16    26 p     comp~
##  6 audi         a4           2.8  1999     6 manu~ f        18    26 p     comp~
```

```
##  7 audi         a4           3.1  2008     6 auto~ f        18    27 p     comp~
##  8 audi         a4 quattro   1.8  1999     4 manu~ 4        18    26 p     comp~
##  9 audi         a4 quattro   1.8  1999     4 auto~ 4        16    25 p     comp~
## 10 audi         a4 quattro   2    2008     4 manu~ 4        20    28 p     comp~
## # i 224 more rows
```

1.2.1 filter, 데이터에서 특정 열의 값이 조건에 맞는 행을 선택할 수 있습니다.

```
filter(mpg, manufacturer=="hyundai")
```

```
## # A tibble: 14 x 11
##    manufacturer model   displ  year   cyl trans    drv     cty   hwy fl    class
##    <chr>        <chr>   <dbl> <int> <int> <chr>    <chr> <int> <int> <chr> <chr>
##  1 hyundai      sonata    2.4  1999     4 auto(14) f        18    26 r     mids~
##  2 hyundai      sonata    2.4  1999     4 manual(~ f        18    27 r     mids~
##  3 hyundai      sonata    2.4  2008     4 auto(14) f        21    30 r     mids~
##  4 hyundai      sonata    2.4  2008     4 manual(~ f        21    31 r     mids~
##  5 hyundai      sonata    2.5  1999     6 auto(14) f        18    26 r     mids~
##  6 hyundai      sonata    2.5  1999     6 manual(~ f        18    26 r     mids~
##  7 hyundai      sonata    3.3  2008     6 auto(15) f        19    28 r     mids~
##  8 hyundai      tiburon   2    1999     4 auto(14) f        19    26 r     subc~
##  9 hyundai      tiburon   2    1999     4 manual(~ f        19    29 r     subc~
## 10 hyundai      tiburon   2    2008     4 manual(~ f        20    28 r     subc~
## 11 hyundai      tiburon   2    2008     4 auto(14) f        20    27 r     subc~
## 12 hyundai      tiburon   2.7  2008     6 auto(14) f        17    24 r     subc~
## 13 hyundai      tiburon   2.7  2008     6 manual(~ f        16    24 r     subc~
## 14 hyundai      tiburon   2.7  2008     6 manual(~ f        17    24 r     subc~
```

```
filter(mpg, cty > 28)
```

```
## # A tibble: 3 x 11
##   manufacturer model      displ  year   cyl trans  drv     cty   hwy fl    class
##   <chr>        <chr>      <dbl> <int> <int> <chr>  <chr> <int> <int> <chr> <chr>
## 1 volkswagen   jetta        1.9  1999     4 manua~ f        33    44 d     comp~
## 2 volkswagen   new beetle   1.9  1999     4 manua~ f        35    44 d     subc~
## 3 volkswagen   new beetle   1.9  1999     4 auto(~ f        29    41 d     subc~
```

```
filter(mpg, cty * 2 > 60) # 연산 후 결과를 기반으로 비교 가능
```

```
## # A tibble: 2 x 11
##   manufacturer model      displ  year   cyl trans  drv     cty   hwy fl    class
##   <chr>        <chr>      <dbl> <int> <int> <chr>  <chr> <int> <int> <chr> <chr>
## 1 volkswagen   jetta        1.9  1999     4 manua~ f        33    44 d     comp~
## 2 volkswagen   new beetle   1.9  1999     4 manua~ f        35    44 d     subc~
```

```
filter(mpg, manufacturer=="hyundai", cty >= 20) # 여러 조건을 한번에 선택 가능
```

```
## # A tibble: 4 x 11
##   manufacturer model   displ  year   cyl trans     drv     cty   hwy fl    class
##   <chr>        <chr>   <dbl> <int> <int> <chr>     <chr> <int> <int> <chr> <chr>
## 1 hyundai      sonata    2.4  2008     4 auto(14)  f        21    30 r     mids~
## 2 hyundai      sonata    2.4  2008     4 manual(m~ f        21    31 r     mids~
## 3 hyundai      tiburon   2    2008     4 manual(m~ f        20    28 r     subc~
## 4 hyundai      tiburon   2    2008     4 auto(14)  f        20    27 r     subc~
```

```
filter(mpg, model=="sonata" | cty >= 28) # 논리 연산자를 사용가능
```

```
## # A tibble: 12 x 11
##    manufacturer model      displ  year   cyl trans drv     cty   hwy fl    class
##    <chr>        <chr>      <dbl> <int> <int> <chr> <chr> <int> <int> <chr> <chr>
##  1 honda        civic        1.6  1999     4 manu~ f        28    33 r     subc~
##  2 hyundai      sonata       2.4  1999     4 auto~ f        18    26 r     mids~
##  3 hyundai      sonata       2.4  1999     4 manu~ f        18    27 r     mids~
##  4 hyundai      sonata       2.4  2008     4 auto~ f        21    30 r     mids~
##  5 hyundai      sonata       2.4  2008     4 manu~ f        21    31 r     mids~
##  6 hyundai      sonata       2.5  1999     6 auto~ f        18    26 r     mids~
##  7 hyundai      sonata       2.5  1999     6 manu~ f        18    26 r     mids~
##  8 hyundai      sonata       3.3  2008     6 auto~ f        19    28 r     mids~
##  9 toyota       corolla      1.8  2008     4 manu~ f        28    37 r     comp~
## 10 volkswagen   jetta        1.9  1999     4 manu~ f        33    44 d     comp~
## 11 volkswagen   new beetle   1.9  1999     4 manu~ f        35    44 d     subc~
## 12 volkswagen   new beetle   1.9  1999     4 auto~ f        29    41 d     subc~
```

```
filter(mpg, model=="sonata" | cty >= 28, year==2008)
```

```
## # A tibble: 4 x 11
##   manufacturer model   displ  year   cyl trans     drv     cty   hwy fl    class
##   <chr>        <chr>   <dbl> <int> <int> <chr>     <chr> <int> <int> <chr> <chr>
## 1 hyundai      sonata    2.4  2008     4 auto(14)  f        21    30 r     mids~
## 2 hyundai      sonata    2.4  2008     4 manual(m~ f        21    31 r     mids~
## 3 hyundai      sonata    3.3  2008     6 auto(15)  f        19    28 r     mids~
## 4 toyota       corolla   1.8  2008     4 manual(m~ f        28    37 r     comp~
```

%in% 연산자를 사용하면 다양한 형태의 조건식을 활용할 수 있습니다.

```
filter(mpg, year==2008, hwy >= 28, model %in% c("sonata","corolla","jetta"))
```

```
## # A tibble: 9 x 11
##   manufacturer model   displ  year   cyl trans     drv     cty   hwy fl    class
##   <chr>        <chr>   <dbl> <int> <int> <chr>     <chr> <int> <int> <chr> <chr>
## 1 hyundai      sonata    2.4  2008     4 auto(14)  f        21    30 r     mids~
## 2 hyundai      sonata    2.4  2008     4 manual(m~ f        21    31 r     mids~
## 3 hyundai      sonata    3.3  2008     6 auto(15)  f        19    28 r     mids~
## 4 toyota       corolla   1.8  2008     4 manual(m~ f        28    37 r     comp~
## 5 toyota       corolla   1.8  2008     4 auto(14)  f        26    35 r     comp~
## 6 volkswagen   jetta     2    2008     4 auto(s6)  f        22    29 p     comp~
## 7 volkswagen   jetta     2    2008     4 manual(m~ f        21    29 p     comp~
## 8 volkswagen   jetta     2.5  2008     5 auto(s6)  f        21    29 r     comp~
## 9 volkswagen   jetta     2.5  2008     5 manual(m~ f        21    29 r     comp~
```

1.2.2 slice, 데이터에서 특정 위치의 행을 선택

slice를 사용하면 특정 위치의 행을 선택할 수 있습니다.

```
hyundai_2008 <- filter(mpg, manufacturer == "hyundai", year == 2008)
slice(hyundai_2008, 1) # 1 행 선택
```

```
## # A tibble: 1 x 11
##   manufacturer model  displ  year   cyl trans    drv     cty   hwy fl    class
##   <chr>        <chr>  <dbl> <int> <int> <chr>    <chr> <int> <int> <chr> <chr>
## 1 hyundai      sonata   2.4  2008     4 auto(14) f        21    30 r     midsize
```

```r
slice(hyundai_2008, 1:3) # 1-3 행 선택
```

```
## # A tibble: 3 x 11
##   manufacturer model  displ  year   cyl trans      drv     cty   hwy fl    class
##   <chr>        <chr>  <dbl> <int> <int> <chr>      <chr> <int> <int> <chr> <chr>
## 1 hyundai      sonata   2.4  2008     4 auto(l4)   f        21    30 r     mids~
## 2 hyundai      sonata   2.4  2008     4 manual(m5) f        21    31 r     mids~
## 3 hyundai      sonata   3.3  2008     6 auto(l5)   f        19    28 r     mids~
```

```r
slice(hyundai_2008, 1:3, 6:7) # 1-3과 6-7 행 선택
```

```
## # A tibble: 5 x 11
##   manufacturer model   displ  year   cyl trans     drv     cty   hwy fl    class
##   <chr>        <chr>   <dbl> <int> <int> <chr>     <chr> <int> <int> <chr> <chr>
## 1 hyundai      sonata    2.4  2008     4 auto(l4)  f        21    30 r     mids~
## 2 hyundai      sonata    2.4  2008     4 manual(m~ f        21    31 r     mids~
## 3 hyundai      sonata    3.3  2008     6 auto(l5)  f        19    28 r     mids~
## 4 hyundai      tiburon   2.7  2008     6 auto(l4)  f        17    24 r     subc~
## 5 hyundai      tiburon   2.7  2008     6 manual(m~ f        16    24 r     subc~
```

```r
slice(hyundai_2008, -1) # 1행 제외
```

```
## # A tibble: 7 x 11
##   manufacturer model   displ  year   cyl trans     drv     cty   hwy fl    class
##   <chr>        <chr>   <dbl> <int> <int> <chr>     <chr> <int> <int> <chr> <chr>
## 1 hyundai      sonata    2.4  2008     4 manual(m~ f        21    31 r     mids~
## 2 hyundai      sonata    3.3  2008     6 auto(l5)  f        19    28 r     mids~
## 3 hyundai      tiburon   2    2008     4 manual(m~ f        20    28 r     subc~
## 4 hyundai      tiburon   2    2008     4 auto(l4)  f        20    27 r     subc~
## 5 hyundai      tiburon   2.7  2008     6 auto(l4)  f        17    24 r     subc~
## 6 hyundai      tiburon   2.7  2008     6 manual(m~ f        16    24 r     subc~
## 7 hyundai      tiburon   2.7  2008     6 manual(m~ f        17    24 r     subc~
```

```r
slice(hyundai_2008, -1, -(4:6))  # 1 행과 4-6 행 제외
```

```
## # A tibble: 4 x 11
##   manufacturer model   displ  year   cyl trans     drv     cty   hwy fl    class
##   <chr>        <chr>   <dbl> <int> <int> <chr>     <chr> <int> <int> <chr> <chr>
## 1 hyundai      sonata    2.4  2008     4 manual(m~ f        21    31 r     mids~
## 2 hyundai      sonata    3.3  2008     6 auto(l5)  f        19    28 r     mids~
## 3 hyundai      tiburon   2.7  2008     6 manual(m~ f        16    24 r     subc~
## 4 hyundai      tiburon   2.7  2008     6 manual(m~ f        17    24 r     subc~
```

slice_sample을 사용하면 데이터에서 원하는 수 또는 비율만큼 행을 임의 추출할 수 있습니다. slince_로 시작하는 다양한 선택 함수를 제공합니다.

```r
slice_sample(hyundai_2008, n=10)
```

```
## # A tibble: 8 x 11
##   manufacturer model   displ  year   cyl trans     drv     cty   hwy fl    class
##   <chr>        <chr>   <dbl> <int> <int> <chr>     <chr> <int> <int> <chr> <chr>
## 1 hyundai      tiburon   2    2008     4 manual(m~ f        20    28 r     subc~
## 2 hyundai      tiburon   2.7  2008     6 auto(l4)  f        17    24 r     subc~
## 3 hyundai      sonata    3.3  2008     6 auto(l5)  f        19    28 r     mids~
## 4 hyundai      tiburon   2.7  2008     6 manual(m~ f        16    24 r     subc~
## 5 hyundai      tiburon   2.7  2008     6 manual(m~ f        17    24 r     subc~
## 6 hyundai      sonata    2.4  2008     4 manual(m~ f        21    31 r     mids~
```

```
## 7 hyundai      tiburon   2    2008     4 auto(l4)  f        20    27 r     subc~
## 8 hyundai      sonata    2.4  2008     4 auto(l4)  f        21    30 r     mids~
```

```
slice_sample(hyundai_2008, prop=0.8) # 데이터에서 80% 행을 추출
```

```
## # A tibble: 6 x 11
##   manufacturer model   displ  year   cyl trans     drv     cty   hwy fl    class
##   <chr>        <chr>   <dbl> <int> <int> <chr>     <chr> <int> <int> <chr> <chr>
## 1 hyundai      tiburon   2    2008     4 auto(l4)  f        20    27 r     subc~
## 2 hyundai      sonata    2.4  2008     4 auto(l4)  f        21    30 r     mids~
## 3 hyundai      sonata    2.4  2008     4 manual(m~ f        21    31 r     mids~
## 4 hyundai      tiburon   2.7  2008     6 manual(m~ f        16    24 r     subc~
## 5 hyundai      sonata    3.3  2008     6 auto(l5)  f        19    28 r     mids~
## 6 hyundai      tiburon   2.7  2008     6 auto(l4)  f        17    24 r     subc~
```

```
slice_head(mpg, n=4)  # 데이터의 처음 4 행 추출
```

```
## # A tibble: 4 x 11
##   manufacturer model displ  year   cyl trans      drv     cty   hwy fl    class
##   <chr>        <chr> <dbl> <int> <int> <chr>      <chr> <int> <int> <chr> <chr>
## 1 audi         a4      1.8  1999     4 auto(l5)   f        18    29 p     compa~
## 2 audi         a4      1.8  1999     4 manual(m5) f        21    29 p     compa~
## 3 audi         a4      2    2008     4 manual(m6) f        20    31 p     compa~
## 4 audi         a4      2    2008     4 auto(av)   f        21    30 p     compa~
```

```
slice_tail(mpg, prop=0.05) # 데이터의 마지막 5% 추출
```

```
## # A tibble: 11 x 11
##    manufacturer model      displ  year   cyl trans drv     cty   hwy fl    class
##    <chr>        <chr>      <dbl> <int> <int> <chr> <chr> <int> <int> <chr> <chr>
##  1 volkswagen   new beetle   2    1999     4 manu~ f        21    29 r     subc~
##  2 volkswagen   new beetle   2    1999     4 auto~ f        19    26 r     subc~
##  3 volkswagen   new beetle   2.5  2008     5 manu~ f        20    28 r     subc~
##  4 volkswagen   new beetle   2.5  2008     5 auto~ f        20    29 r     subc~
##  5 volkswagen   passat       1.8  1999     4 manu~ f        21    29 p     mids~
##  6 volkswagen   passat       1.8  1999     4 auto~ f        18    29 p     mids~
##  7 volkswagen   passat       2    2008     4 auto~ f        19    28 p     mids~
##  8 volkswagen   passat       2    2008     4 manu~ f        21    29 p     mids~
##  9 volkswagen   passat       2.8  1999     6 auto~ f        16    26 p     mids~
## 10 volkswagen   passat       2.8  1999     6 manu~ f        18    26 p     mids~
## 11 volkswagen   passat       3.6  2008     6 auto~ f        17    26 p     mids~
```

```
slice_min(mpg, cty, n=2) # cty 열의 값이 가장 작은 2 행 추출
```

```
## # A tibble: 5 x 11
##   manufacturer model       displ  year   cyl trans drv     cty   hwy fl    class
##   <chr>        <chr>       <dbl> <int> <int> <chr> <chr> <int> <int> <chr> <chr>
## 1 dodge        dakota pic~   4.7  2008     8 auto~ 4         9    12 e     pick~
## 2 dodge        durango 4wd   4.7  2008     8 auto~ 4         9    12 e     suv
## 3 dodge        ram 1500 p~   4.7  2008     8 auto~ 4         9    12 e     pick~
## 4 dodge        ram 1500 p~   4.7  2008     8 manu~ 4         9    12 e     pick~
## 5 jeep         grand cher~   4.7  2008     8 auto~ 4         9    12 e     suv
```

```
slice_min(mpg, cty, n=2, with_ties = F) # 동률 행을 추출하지 않는다.
```

```
## # A tibble: 2 x 11
##   manufacturer model       displ  year   cyl trans drv     cty   hwy fl    class
```

```
##   <chr>        <chr>       <dbl> <int> <int> <chr> <chr> <int> <int> <chr> <chr>
## 1 dodge        dakota pic~   4.7  2008     8 auto~ 4         9    12 e     pick~
## 2 dodge        durango 4wd   4.7  2008     8 auto~ 4         9    12 e     suv
```

1.2.3 arrange, 특정 열의 값을 기준으로 데이터의 행을 정렬

```
arrange(hyundai_2008, cyl)
```

```
## # A tibble: 8 x 11
##   manufacturer model   displ  year   cyl trans     drv     cty   hwy fl    class
##   <chr>        <chr>   <dbl> <int> <int> <chr>     <chr> <int> <int> <chr> <chr>
## 1 hyundai      sonata    2.4  2008     4 auto(14)  f        21    30 r     mids~
## 2 hyundai      sonata    2.4  2008     4 manual(m~ f        21    31 r     mids~
## 3 hyundai      tiburon   2    2008     4 manual(m~ f        20    28 r     subc~
## 4 hyundai      tiburon   2    2008     4 auto(14)  f        20    27 r     subc~
## 5 hyundai      sonata    3.3  2008     6 auto(15)  f        19    28 r     mids~
## 6 hyundai      tiburon   2.7  2008     6 auto(14)  f        17    24 r     subc~
## 7 hyundai      tiburon   2.7  2008     6 manual(m~ f        16    24 r     subc~
## 8 hyundai      tiburon   2.7  2008     6 manual(m~ f        17    24 r     subc~
```

```
arrange(hyundai_2008, cyl, cty)
```

```
## # A tibble: 8 x 11
##   manufacturer model   displ  year   cyl trans     drv     cty   hwy fl    class
##   <chr>        <chr>   <dbl> <int> <int> <chr>     <chr> <int> <int> <chr> <chr>
## 1 hyundai      tiburon   2    2008     4 manual(m~ f        20    28 r     subc~
## 2 hyundai      tiburon   2    2008     4 auto(14)  f        20    27 r     subc~
## 3 hyundai      sonata    2.4  2008     4 auto(14)  f        21    30 r     mids~
## 4 hyundai      sonata    2.4  2008     4 manual(m~ f        21    31 r     mids~
## 5 hyundai      tiburon   2.7  2008     6 manual(m~ f        16    24 r     subc~
## 6 hyundai      tiburon   2.7  2008     6 auto(14)  f        17    24 r     subc~
## 7 hyundai      tiburon   2.7  2008     6 manual(m~ f        17    24 r     subc~
## 8 hyundai      sonata    3.3  2008     6 auto(15)  f        19    28 r     mids~
```

```
arrange(hyundai_2008, model, trans)
```

```
## # A tibble: 8 x 11
##   manufacturer model   displ  year   cyl trans     drv     cty   hwy fl    class
##   <chr>        <chr>   <dbl> <int> <int> <chr>     <chr> <int> <int> <chr> <chr>
## 1 hyundai      sonata    2.4  2008     4 auto(14)  f        21    30 r     mids~
## 2 hyundai      sonata    3.3  2008     6 auto(15)  f        19    28 r     mids~
## 3 hyundai      sonata    2.4  2008     4 manual(m~ f        21    31 r     mids~
## 4 hyundai      tiburon   2    2008     4 auto(14)  f        20    27 r     subc~
## 5 hyundai      tiburon   2.7  2008     6 auto(14)  f        17    24 r     subc~
## 6 hyundai      tiburon   2    2008     4 manual(m~ f        20    28 r     subc~
## 7 hyundai      tiburon   2.7  2008     6 manual(m~ f        17    24 r     subc~
## 8 hyundai      tiburon   2.7  2008     6 manual(m~ f        16    24 r     subc~
```

```
arrange(hyundai_2008, desc(cyl))
```

```
## # A tibble: 8 x 11
##   manufacturer model   displ  year   cyl trans     drv     cty   hwy fl    class
##   <chr>        <chr>   <dbl> <int> <int> <chr>     <chr> <int> <int> <chr> <chr>
## 1 hyundai      sonata    3.3  2008     6 auto(15)  f        19    28 r     mids~
## 2 hyundai      tiburon   2.7  2008     6 auto(14)  f        17    24 r     subc~
## 3 hyundai      tiburon   2.7  2008     6 manual(m~ f        16    24 r     subc~
```

```
## 4 hyundai      tiburon   2.7  2008     6 manual(m~ f        17    24 r     subc~
## 5 hyundai      sonata    2.4  2008     4 auto(l4)  f        21    30 r     mids~
## 6 hyundai      sonata    2.4  2008     4 manual(m~ f        21    31 r     mids~
## 7 hyundai      tiburon   2    2008     4 manual(m~ f        20    28 r     subc~
## 8 hyundai      tiburon   2    2008     4 auto(l4)  f        20    27 r     subc~
```

```
arrange(hyundai_2008, desc(cyl), cty)
```

```
## # A tibble: 8 x 11
##   manufacturer model   displ  year   cyl trans     drv     cty   hwy fl    class
##   <chr>        <chr>   <dbl> <int> <int> <chr>     <chr> <int> <int> <chr> <chr>
## 1 hyundai      tiburon   2.7  2008     6 manual(m~ f        16    24 r     subc~
## 2 hyundai      tiburon   2.7  2008     6 auto(l4)  f        17    24 r     subc~
## 3 hyundai      tiburon   2.7  2008     6 manual(m~ f        17    24 r     subc~
## 4 hyundai      sonata    3.3  2008     6 auto(l5)  f        19    28 r     mids~
## 5 hyundai      tiburon   2    2008     4 manual(m~ f        20    28 r     subc~
## 6 hyundai      tiburon   2    2008     4 auto(l4)  f        20    27 r     subc~
## 7 hyundai      sonata    2.4  2008     4 auto(l4)  f        21    30 r     mids~
## 8 hyundai      sonata    2.4  2008     4 manual(m~ f        21    31 r     mids~
```

```
arrange(hyundai_2008, model, desc(trans))
```

```
## # A tibble: 8 x 11
##   manufacturer model   displ  year   cyl trans     drv     cty   hwy fl    class
##   <chr>        <chr>   <dbl> <int> <int> <chr>     <chr> <int> <int> <chr> <chr>
## 1 hyundai      sonata    2.4  2008     4 manual(m~ f        21    31 r     mids~
## 2 hyundai      sonata    3.3  2008     6 auto(l5)  f        19    28 r     mids~
## 3 hyundai      sonata    2.4  2008     4 auto(l4)  f        21    30 r     mids~
## 4 hyundai      tiburon   2.7  2008     6 manual(m~ f        16    24 r     subc~
## 5 hyundai      tiburon   2    2008     4 manual(m~ f        20    28 r     subc~
## 6 hyundai      tiburon   2.7  2008     6 manual(m~ f        17    24 r     subc~
## 7 hyundai      tiburon   2    2008     4 auto(l4)  f        20    27 r     subc~
## 8 hyundai      tiburon   2.7  2008     6 auto(l4)  f        17    24 r     subc~
```

1.2.4 select, 열의 이름, 위치, 데이터 형식 등으로 열을 선택

```
select(hyundai_2008, model, year, cty, hwy)
```

```
## # A tibble: 8 x 4
##   model    year   cty   hwy
##   <chr>   <int> <int> <int>
## 1 sonata   2008    21    30
## 2 sonata   2008    21    31
## 3 sonata   2008    19    28
## 4 tiburon  2008    20    28
## 5 tiburon  2008    20    27
## 6 tiburon  2008    17    24
## 7 tiburon  2008    16    24
## 8 tiburon  2008    17    24
```

```
select(hyundai_2008, model:trans, cty:hwy)
```

```
## # A tibble: 8 x 7
##   model   displ  year   cyl trans        cty   hwy
##   <chr>   <dbl> <int> <int> <chr>      <int> <int>
## 1 sonata    2.4  2008     4 auto(l4)      21    30
```

```
## 2 sonata    2.4  2008     4 manual(m5)    21    31
## 3 sonata    3.3  2008     6 auto(l5)      19    28
## 4 tiburon   2    2008     4 manual(m5)    20    28
## 5 tiburon   2    2008     4 auto(l4)      20    27
## 6 tiburon   2.7  2008     6 auto(l4)      17    24
## 7 tiburon   2.7  2008     6 manual(m6)    16    24
## 8 tiburon   2.7  2008     6 manual(m5)    17    24
```

```
select(hyundai_2008, -(model:trans))
```

```
## # A tibble: 8 x 6
##   manufacturer drv     cty   hwy fl    class
##   <chr>        <chr> <int> <int> <chr> <chr>
## 1 hyundai      f        21    30 r     midsize
## 2 hyundai      f        21    31 r     midsize
## 3 hyundai      f        19    28 r     midsize
## 4 hyundai      f        20    28 r     subcompact
## 5 hyundai      f        20    27 r     subcompact
## 6 hyundai      f        17    24 r     subcompact
## 7 hyundai      f        16    24 r     subcompact
## 8 hyundai      f        17    24 r     subcompact
```

```
select(hyundai_2008, 1:3) # 1 번부터 3번 열 선택
```

```
## # A tibble: 8 x 3
##   manufacturer model   displ
##   <chr>        <chr>   <dbl>
## 1 hyundai      sonata    2.4
## 2 hyundai      sonata    2.4
## 3 hyundai      sonata    3.3
## 4 hyundai      tiburon   2
## 5 hyundai      tiburon   2
## 6 hyundai      tiburon   2.7
## 7 hyundai      tiburon   2.7
## 8 hyundai      tiburon   2.7
```

```
select(hyundai_2008, 1:3, 5) # 1 번부터 3번 열과 5 번 열 선택
```

```
## # A tibble: 8 x 4
##   manufacturer model   displ   cyl
##   <chr>        <chr>   <dbl> <int>
## 1 hyundai      sonata    2.4     4
## 2 hyundai      sonata    2.4     4
## 3 hyundai      sonata    3.3     6
## 4 hyundai      tiburon   2       4
## 5 hyundai      tiburon   2       4
## 6 hyundai      tiburon   2.7     6
## 7 hyundai      tiburon   2.7     6
## 8 hyundai      tiburon   2.7     6
```

```
select(hyundai_2008, -(4:10)) # 4 번부터 10 번 열 제외하고 선택
```

```
## # A tibble: 8 x 4
##   manufacturer model   displ class
##   <chr>        <chr>   <dbl> <chr>
## 1 hyundai      sonata    2.4 midsize
## 2 hyundai      sonata    2.4 midsize
```

```
## 3 hyundai      sonata    3.3 midsize
## 4 hyundai      tiburon   2   subcompact
## 5 hyundai      tiburon   2   subcompact
## 6 hyundai      tiburon   2.7 subcompact
## 7 hyundai      tiburon   2.7 subcompact
## 8 hyundai      tiburon   2.7 subcompact
```

dplyr 패키지는 다양한 형태로 변수를 선택할 수 있도록 다음의 변수 이름 매칭 함수를 제공합니다. 이러한 함수는 변수의 수가 많을 때 매우 유용합니다.

- starts_with("abs"), abc로 이름이 시작하는 모든 변수
- ends_with("abs"), abc로 이름이 끝나는 모든 변수
- contains("abs"), abc를 이름에 포함하고 있는 모든 변수
- matches("(.)\1"), 정규 표현식을 만족하는 이름을 가진 모든 변수
- num_range("x", 1:3), "x1", "x2", "x3"이라는 이름의 변수

```
select(hyundai_2008, starts_with("c"))
```

```
## # A tibble: 8 x 3
##     cyl   cty class
##   <int> <int> <chr>
## 1     4    21 midsize
## 2     4    21 midsize
## 3     6    19 midsize
## 4     4    20 subcompact
## 5     4    20 subcompact
## 6     6    17 subcompact
## 7     6    16 subcompact
## 8     6    17 subcompact
```

select 함수에 where 함수를 사용하면 해당 조건에 맞는 열만 매칭하여 선택할 수 있습니다. where 함수는 유일한 인수로 함수를 입력받는데, 이 함수는 논리값을 반환하는 함수여야 하며, where 함수는 각 열에 이 함수를 적용합니다. select 함수는 where의 결과가 TRUE인 열만 선택합니다.

where 함수가 주로 사용되는 곳은 데이터 형식에 따라 열을 선택할 때입니다. `is.`로 시작하는 함수들은 어떤 객체가 특정 형식인지를 테스트 합니다. 다음 예는 where 와 is.character 함수를 사용하여 문자열인 열만 선택한 것입니다.

```
select(hyundai_2008, where(is.character))
```

```
## # A tibble: 8 x 6
##   manufacturer model   trans      drv   fl    class
##   <chr>        <chr>   <chr>      <chr> <chr> <chr>
## 1 hyundai      sonata  auto(l4)   f     r     midsize
## 2 hyundai      sonata  manual(m5) f     r     midsize
## 3 hyundai      sonata  auto(l5)   f     r     midsize
## 4 hyundai      tiburon manual(m5) f     r     subcompact
## 5 hyundai      tiburon auto(l4)   f     r     subcompact
## 6 hyundai      tiburon auto(l4)   f     r     subcompact
## 7 hyundai      tiburon manual(m6) f     r     subcompact
## 8 hyundai      tiburon manual(m5) f     r     subcompact
```

```
select(hyundai_2008, where(function(x) is.numeric(x) && mean(x) >= 10))
```

```
## # A tibble: 8 x 3
##    year   cty   hwy
##   <int> <int> <int>
## 1  2008    21    30
```

```
## 2  2008    21    31
## 3  2008    19    28
## 4  2008    20    28
## 5  2008    20    27
## 6  2008    17    24
## 7  2008    16    24
## 8  2008    17    24
```

```
select(hyundai_2008, where(~ is.numeric(.x) && mean(.x) < 10)) # purrr 형식
```

```
## # A tibble: 8 x 2
##   displ   cyl
##   <dbl> <int>
## 1   2.4     4
## 2   2.4     4
## 3   3.3     6
## 4   2       4
## 5   2       4
## 6   2.7     6
## 7   2.7     6
## 8   2.7     6
```

select 함수에서 변수 이름을 지정할 때, (새로운 변수 이름)=(기존 변수 이름) 형식으로 지정하면 변수의 이름을 바꿀 수 있습니다.

```
select(hyundai_2008, model, city=cty, highway=hwy)
```

```
## # A tibble: 8 x 3
##   model    city highway
##   <chr>   <int>   <int>
## 1 sonata     21      30
## 2 sonata     21      31
## 3 sonata     19      28
## 4 tiburon    20      28
## 5 tiburon    20      27
## 6 tiburon    17      24
## 7 tiburon    16      24
## 8 tiburon    17      24
```

```
rename(hyundai_2008, city=cty, highway=hwy) # 변수 이름만 변경 가능
```

```
## # A tibble: 8 x 11
##   manufacturer model   displ  year   cyl trans   drv    city highway fl    class
##   <chr>        <chr>   <dbl> <int> <int> <chr>   <chr> <int>   <int> <chr> <chr>
## 1 hyundai      sonata    2.4  2008     4 auto(l~ f        21      30 r     mids~
## 2 hyundai      sonata    2.4  2008     4 manual~ f        21      31 r     mids~
## 3 hyundai      sonata    3.3  2008     6 auto(l~ f        19      28 r     mids~
## 4 hyundai      tiburon   2    2008     4 manual~ f        20      28 r     subc~
## 5 hyundai      tiburon   2    2008     4 auto(l~ f        20      27 r     subc~
## 6 hyundai      tiburon   2.7  2008     6 auto(l~ f        17      24 r     subc~
## 7 hyundai      tiburon   2.7  2008     6 manual~ f        16      24 r     subc~
## 8 hyundai      tiburon   2.7  2008     6 manual~ f        17      24 r     subc~
```

select 함수는 나열된 변수의 순서에 따라 새롭게 만들어진 데이터 프레임의 변수의 순서를 조정 가능합니다.

```
select(hyundai_2008, cty, hwy)
```

```
## # A tibble: 8 x 2
```

```
##     cty   hwy
##   <int> <int>
## 1    21    30
## 2    21    31
## 3    19    28
## 4    20    28
## 5    20    27
## 6    17    24
## 7    16    24
## 8    17    24
```

```
select(hyundai_2008, hwy, cty)
```

```
## # A tibble: 8 x 2
##     hwy   cty
##   <int> <int>
## 1    30    21
## 2    31    21
## 3    28    19
## 4    28    20
## 5    27    20
## 6    24    17
## 7    24    16
## 8    24    17
```

```
select(hyundai_2008, cty, hwy, everything())
```

```
## # A tibble: 8 x 11
##     cty   hwy manufacturer model   displ  year   cyl trans     drv   fl    class
##   <int> <int> <chr>        <chr>   <dbl> <int> <int> <chr>     <chr> <chr> <chr>
## 1    21    30 hyundai      sonata    2.4  2008     4 auto(l4)  f     r     mids~
## 2    21    31 hyundai      sonata    2.4  2008     4 manual(m~ f     r     mids~
## 3    19    28 hyundai      sonata    3.3  2008     6 auto(l5)  f     r     mids~
## 4    20    28 hyundai      tiburon   2    2008     4 manual(m~ f     r     subc~
## 5    20    27 hyundai      tiburon   2    2008     4 auto(l4)  f     r     subc~
## 6    17    24 hyundai      tiburon   2.7  2008     6 auto(l4)  f     r     subc~
## 7    16    24 hyundai      tiburon   2.7  2008     6 manual(m~ f     r     subc~
## 8    17    24 hyundai      tiburon   2.7  2008     6 manual(m~ f     r     subc~
```

1.2.5 mutate, 기존 열을 사용하여 새로운 열을 데이터에 추가

mutate는 기존 변수를 이용하여 새로운 변수를 만들어 데이터 프레임의 가장 마지막 열로 추가합니다.

```
hyundai_2008_displ <- select(hyundai_2008, -(cyl:drv), -(fl:class))
mutate(hyundai_2008_displ, sum=cty + hwy)
```

```
## # A tibble: 8 x 7
##   manufacturer model   displ  year   cty   hwy   sum
##   <chr>        <chr>   <dbl> <int> <int> <int> <int>
## 1 hyundai      sonata    2.4  2008    21    30    51
## 2 hyundai      sonata    2.4  2008    21    31    52
## 3 hyundai      sonata    3.3  2008    19    28    47
## 4 hyundai      tiburon   2    2008    20    28    48
## 5 hyundai      tiburon   2    2008    20    27    47
## 6 hyundai      tiburon   2.7  2008    17    24    41
## 7 hyundai      tiburon   2.7  2008    16    24    40
```

```
## 8 hyundai      tiburon   2.7  2008    17    24    41
```

```
mutate(hyundai_2008_displ,
       sum=cty + hwy,
       mean=(cty + hwy) / 2,
       ratio= cty / hwy * 100)
```

```
## # A tibble: 8 x 9
##   manufacturer model   displ  year   cty   hwy   sum  mean ratio
##   <chr>        <chr>   <dbl> <int> <int> <int> <int> <dbl> <dbl>
## 1 hyundai      sonata    2.4  2008    21    30    51  25.5  70
## 2 hyundai      sonata    2.4  2008    21    31    52  26    67.7
## 3 hyundai      sonata    3.3  2008    19    28    47  23.5  67.9
## 4 hyundai      tiburon   2    2008    20    28    48  24    71.4
## 5 hyundai      tiburon   2    2008    20    27    47  23.5  74.1
## 6 hyundai      tiburon   2.7  2008    17    24    41  20.5  70.8
## 7 hyundai      tiburon   2.7  2008    16    24    40  20    66.7
## 8 hyundai      tiburon   2.7  2008    17    24    41  20.5  70.8
```

만약 새롭게 만들어진 변수만 데이터에 남기려면 mutate() 대신 transmute()를 사용합니다.

```
transmute(hyundai_2008_displ,
          sum=cty + hwy,
          mean=(cty + hwy) / 2,
          ratio= cty / hwy * 100)
```

```
## # A tibble: 8 x 3
##     sum  mean ratio
##   <int> <dbl> <dbl>
## 1    51  25.5  70
## 2    52  26    67.7
## 3    47  23.5  67.9
## 4    48  24    71.4
## 5    47  23.5  74.1
## 6    41  20.5  70.8
## 7    40  20    66.7
## 8    41  20.5  70.8
```

새로운 변수를 생성할 때, 기존 변수와 관련된 수치, 논리, 문자열 연산을 수행할 수 있습니다. 다음처럼 제조사와 모델을 하나로 합쳐서 새로운 변수를 만들수도 있고, 배기량이 3 이상인지 여부를 나타내는 변수도 만들 수 있습니다.

```
mutate(hyundai_2008_displ, newName=paste(manufacturer, model, sep="-"), dis3=displ >= 3)
```

```
## # A tibble: 8 x 8
##   manufacturer model   displ  year   cty   hwy newName         dis3
##   <chr>        <chr>   <dbl> <int> <int> <int> <chr>           <lgl>
## 1 hyundai      sonata    2.4  2008    21    30 hyundai-sonata  FALSE
## 2 hyundai      sonata    2.4  2008    21    31 hyundai-sonata  FALSE
## 3 hyundai      sonata    3.3  2008    19    28 hyundai-sonata  TRUE
## 4 hyundai      tiburon   2    2008    20    28 hyundai-tiburon FALSE
## 5 hyundai      tiburon   2    2008    20    27 hyundai-tiburon FALSE
## 6 hyundai      tiburon   2.7  2008    17    24 hyundai-tiburon FALSE
## 7 hyundai      tiburon   2.7  2008    16    24 hyundai-tiburon FALSE
## 8 hyundai      tiburon   2.7  2008    17    24 hyundai-tiburon FALSE
```

다음 함수가 새로운 변수를 만들 때 자주 사용됩니다.

- row_number(), 각 행의 행번호, 각 행에 일련번호를 붙일 때 유용
- lead(), 기존 변수를 한 행, 또는 여러 행 빠르게 시작하는 변수
- lag(), 기존 변수를 한 행, 또는 여러 행 늦게 시작하는 변수
- cumsum()/cummean(), 누적 합과 평균
- min_rank(), 가장 작은 것부터 차례대로 크기 순서로 등수를 매기는 함수, desc() 함수를 변수에 적용한 후 등수를 매기면 가장 큰 것부터 순서를 매길 수 있음
- dense_rank(), percent_rank(), cume_dist(), ntile() 등

```
store <- data.frame(month=1:6, sales=c(23, 45, 34, 67, 30, 41))
store
```

```
##   month sales
## 1     1    23
## 2     2    45
## 3     3    34
## 4     4    67
## 5     5    30
## 6     6    41
```

```
mutate(store,
       before = lag(sales, n = 1),    # 1 달 전 판매량
       after = lead(sales, n = 1),    # 1 달 후 판매량
       total = cumsum(sales),         # 누적 판매량
       mean = cummean(sales),         # 누적 평균 판매량
       rank1 = min_rank(sales),    # 판매량 순위 (올림차순)
       rank2 = min_rank(desc(sales))    # 판매량 순위 (내림차순)
       )
```

```
##   month sales before after total  mean rank1 rank2
## 1     1    23     NA    45    23 23.00     1     6
## 2     2    45     23    34    68 34.00     5     2
## 3     3    34     45    67   102 34.00     3     4
## 4     4    67     34    30   169 42.25     6     1
## 5     5    30     67    41   199 39.80     2     5
## 6     6    41     30    NA   240 40.00     4     3
```

```
mutate(hyundai_2008_displ, id = row_number())
```

```
## # A tibble: 8 x 7
##   manufacturer model   displ  year   cty   hwy    id
##   <chr>        <chr>   <dbl> <int> <int> <int> <int>
## 1 hyundai      sonata    2.4  2008    21    30     1
## 2 hyundai      sonata    2.4  2008    21    31     2
## 3 hyundai      sonata    3.3  2008    19    28     3
## 4 hyundai      tiburon   2    2008    20    28     4
## 5 hyundai      tiburon   2    2008    20    27     5
## 6 hyundai      tiburon   2.7  2008    17    24     6
## 7 hyundai      tiburon   2.7  2008    16    24     7
## 8 hyundai      tiburon   2.7  2008    17    24     8
```

1.2.6 summarize, 데이터 전체 또는 특정 열을 하나의 통계량으로 요약

summarize 함수는 데이터프레임을 하나의 행으로 요약합니다. 하나의 행으로 요약하기 위하여 변수의 모든 값을 하나의 값으로 통계요약하는 함수를 주로 이용하는데, 대표적인 통계요약 함수는 다음과 같습니다.

- n(), 행의 수
- sum(), 수치 변수의 합

- mean(), 수치 변수의 균
- median(), 수치 변수의 중위수
- sd(), 수치 변수의 표준편차
- var(), 수치 변수의 분산
- min(), 수치 변수의 최소값
- max(), 수치 변수의 최대값
- quantile(변수, probs), 수치 변수의probs‘ 분위수

```
summarize(hyundai_2008_displ,
          count=n(),
          mean=mean(cty),
          med=median(cty),
          min=min(cty),
          max=max(cty))
```

```
## # A tibble: 1 x 5
##   count  mean   med   min   max
##   <int> <dbl> <dbl> <int> <int>
## 1     8  18.9  19.5    16    21
```

```
summarize(hyundai_2008_displ,
          meanCty=mean(cty),
          meanHwy=mean(hwy),
          medianCty=median(cty),
          medianHqy=median(hwy))
```

```
## # A tibble: 1 x 4
##   meanCty meanHwy medianCty medianHqy
##     <dbl>   <dbl>     <dbl>     <dbl>
## 1    18.9      27      19.5      27.5
```

across 함수는 select 처럼 열을 선택하여 동일한 함수를 적용할 수 있습니다. across 함수는 두 개의 주요 인수를 가지고 있는데, 첫 번째 인수는 .cols로 함수를 적용할 열을 지정하며, 두 번재 인수는 .fns로 열에 적용할 함수를 지정합니다.

```
summarize(hyundai_2008_displ, across(c(cty, hwy), mean))
```

```
## # A tibble: 1 x 2
##     cty   hwy
##   <dbl> <dbl>
## 1  18.9    27
```

```
summarize(hyundai_2008_displ, across(c(cty, hwy), list(mean=mean, med=median)))
```

```
## # A tibble: 1 x 4
##   cty_mean cty_med hwy_mean hwy_med
##      <dbl>   <dbl>    <dbl>   <dbl>
## 1     18.9    19.5       27    27.5
```

```
summarize(hyundai_2008_displ, across(c(cty, hwy), list(mean=mean, med=median), .names="{.fn}-{.col}"))
```

```
## # A tibble: 1 x 4
##   `mean-cty` `med-cty` `mean-hwy` `med-hwy`
##        <dbl>     <dbl>      <dbl>     <dbl>
## 1       18.9      19.5         27      27.5
```

```
summarize(hyundai_2008_displ, across(c(cty, hwy), mean, .names="mean-{.col}"),
                              across(c(cty, hwy), median, .names="med-{.col}"))
```

```
## # A tibble: 1 x 4
##   `mean-cty` `mean-hwy` `med-cty` `med-hwy`
##        <dbl>      <dbl>     <dbl>     <dbl>
## 1       18.9         27      19.5      27.5
```

```
summarize(hyundai_2008_displ, across(where(is.numeric), sd))
```

```
## # A tibble: 1 x 4
##   displ  year   cty   hwy
##   <dbl> <dbl> <dbl> <dbl>
## 1 0.427     0  1.96  2.78
```

1.2.7 group_by(), 데이터 요약에만 사용되는 것은 아니나, 그룹별로 데이터를 통계 요약할 때 자주 사용

```
byModel <- group_by(hyundai_2008_displ, model)
summarize(byModel, count=n(), mean=mean(cty), sd=sd(cty))
```

```
## # A tibble: 2 x 4
##   model   count  mean    sd
##   <chr>   <int> <dbl> <dbl>
## 1 sonata      3  20.3  1.15
## 2 tiburon     5  18    1.87
```

```
byModel <- group_by(hyundai_2008_displ, model, cty)
summarize(byModel, count=n(), mean=mean(cty))
```

```
## `summarise()` has grouped output by 'model'. You can override using the
## `.groups` argument.
```

```
## # A tibble: 5 x 4
## # Groups:   model [2]
##   model     cty count  mean
##   <chr>   <int> <int> <dbl>
## 1 sonata     19     1    19
## 2 sonata     21     2    21
## 3 tiburon    16     1    16
## 4 tiburon    17     2    17
## 5 tiburon    20     2    20
```

group_by 는 summarize 와 함께 자주 사용되지만, mutate()나 filter() 등의 다른 dplyr 함수와 함께 사용될 수 있습니다.

```
mutate(hyundai_2008_displ, rank=min_rank(desc(hwy)))  # 전체 데이터에서 고속도로 연비 순위 매기기
```

```
## # A tibble: 8 x 7
##   manufacturer model   displ  year   cty   hwy  rank
##   <chr>        <chr>   <dbl> <int> <int> <int> <int>
## 1 hyundai      sonata    2.4  2008    21    30     2
## 2 hyundai      sonata    2.4  2008    21    31     1
## 3 hyundai      sonata    3.3  2008    19    28     3
## 4 hyundai      tiburon   2    2008    20    28     3
## 5 hyundai      tiburon   2    2008    20    27     5
## 6 hyundai      tiburon   2.7  2008    17    24     6
## 7 hyundai      tiburon   2.7  2008    16    24     6
## 8 hyundai      tiburon   2.7  2008    17    24     6
```

```
mutate(byModel, rank=min_rank(desc(hwy)))  # 모델 별로 고속도로 연비 순위 매기기
```

```
## # A tibble: 8 x 7
## # Groups:   model, cty [5]
##   manufacturer model   displ  year   cty   hwy  rank
##   <chr>        <chr>   <dbl> <int> <int> <int> <int>
## 1 hyundai      sonata    2.4  2008    21    30     2
## 2 hyundai      sonata    2.4  2008    21    31     1
## 3 hyundai      sonata    3.3  2008    19    28     1
## 4 hyundai      tiburon   2    2008    20    28     1
## 5 hyundai      tiburon   2    2008    20    27     2
## 6 hyundai      tiburon   2.7  2008    17    24     1
## 7 hyundai      tiburon   2.7  2008    16    24     1
## 8 hyundai      tiburon   2.7  2008    17    24     1
```

앞의 예에서 모델 별로 데이터 개수만을 세려면 다음과 같은 명령어를 사용하면 됩니다.

```
byModel <- group_by(hyundai_2008_displ, model)
summarise(byModel, n=n())
```

```
## # A tibble: 2 x 2
##   model       n
##   <chr>   <int>
## 1 sonata      3
## 2 tiburon     5
```

```
count(hyundai_2008_displ, model)
```

```
## # A tibble: 2 x 2
##   model       n
##   <chr>   <int>
## 1 sonata      3
## 2 tiburon     5
```

```
count(mpg, class)
```

```
## # A tibble: 7 x 2
##   class          n
##   <chr>      <int>
## 1 2seater        5
## 2 compact       47
## 3 midsize       41
## 4 minivan       11
## 5 pickup        33
## 6 subcompact    35
## 7 suv           62
```

```
count(mpg, class, sort=TRUE)
```

```
## # A tibble: 7 x 2
##   class          n
##   <chr>      <int>
## 1 suv           62
## 2 compact       47
## 3 midsize       41
## 4 subcompact    35
## 5 pickup        33
```

```
## 6 minivan       11
## 7 2seater        5
```

```
count(mpg, class, cyl, sort=T)
```

```
## # A tibble: 19 x 3
##    class        cyl     n
##    <chr>      <int> <int>
##  1 suv            8    38
##  2 compact        4    32
##  3 midsize        6    23
##  4 subcompact     4    21
##  5 pickup         8    20
##  6 midsize        4    16
##  7 suv            6    16
##  8 compact        6    13
##  9 minivan        6    10
## 10 pickup         6    10
## 11 suv            4     8
## 12 subcompact     6     7
## 13 2seater        8     5
## 14 subcompact     8     5
## 15 pickup         4     3
## 16 compact        5     2
## 17 midsize        8     2
## 18 subcompact     5     2
## 19 minivan        4     1
```

1.2.8 %>% 연산자

mpg 데이터에서 조사 연도와 모델 별로 데이터 수와 도심 연비의 평균을 구한 후, 평균이 22 이상인 모델로 이루어진 행을 추출하려고 합니다. 이를 수행하려면 다음처럼 변수를 이용하여 결과를 차례로 전달하거나, 함수를 결합하여 한 문장에 사용해야 합니다.

```
# 1. 중간 결과를 요구
byModel <- group_by(mpg, model, year)
meanCty <- summarize(byModel, count=n(), mean=mean(cty))
```

```
## `summarise()` has grouped output by 'model'. You can override using the
## `.groups` argument.
```

```
filter(meanCty, mean >= 22)
```

```
## # A tibble: 5 x 4
## # Groups:   model [3]
##   model       year count  mean
##   <chr>      <int> <int> <dbl>
## 1 civic       1999     5  24.8
## 2 civic       2008     4  24
## 3 corolla     1999     3  24.7
## 4 corolla     2008     2  27
## 5 new beetle  1999     4  26
```

```
# 2. 결과가 예상이 되지 않음
filter(summarize(group_by(mpg, model, year), count=n(), mean=mean(cty)), mean >= 22)
```

```
## `summarise()` has grouped output by 'model'. You can override using the
```

```
## `.groups` argument.
```

```
## # A tibble: 5 x 4
## # Groups:   model [3]
##   model       year count  mean
##   <chr>      <int> <int> <dbl>
## 1 civic       1999     5  24.8
## 2 civic       2008     4  24
## 3 corolla     1999     3  24.7
## 4 corolla     2008     2  27
## 5 new beetle  1999     4  26
```

파이프 연산자(%>%)는 데이터 변환이 여러 단계를 거칠 때 불필요한 변수의 생성 없이도 함수 간에 중간 데이터를 전달할 수 있게 해 줍니다. 파이프 연산자는 함수의 결과를 뒷 함수의 첫 번째 인수로 전달해 줍니다. 파이프 연산자를 사용할 때는 그러므로 첫 번째 인수는 생략하여 기술합니다. 파이프 연산자로는 magrittr 패키지가 제공하는 %>% 연산자를 사용할 수도 있고, R 4.1.0 버전부터 도입된 기본 기능에 포함된 |> 연산자를 사용할 수도 있다.

```
mpg |> group_by(model, year) |>
  summarize(count=n(), mean=mean(cty), .groups = 'keep') |>
  filter(mean >= 22)
```

```
## # A tibble: 5 x 4
## # Groups:   model, year [5]
##   model       year count  mean
##   <chr>      <int> <int> <dbl>
## 1 civic       1999     5  24.8
## 2 civic       2008     4  24
## 3 corolla     1999     3  24.7
## 4 corolla     2008     2  27
## 5 new beetle  1999     4  26
```