데이터 가공

filter

메소드

함수	기능
filter()	행 추출
select()	열(변수) 추출
arrange()	정렬
mutate()	변수 추가
summarise()	통계치 산정
group_by()	집단별로 나누기
left_join()	데이터 합치기(열)
bind_rows()	데이터 합치기(행)

조건에 맞는 데이터 추출하기

class	english	science
2	98	50
1	97	60
2	86	78
1	98	58
1	80	65
2	89	98

class	english	science
1	97	60
1	98	58
1	80	65

실습 – filter

```
exam <- read.csv("csv_exam.csv")
exam
# exam에서 class가 1인 경우만 추출하여 출력
#단축키 [Ctrl+Shit+M]으로 %>% 기호 입력
exam %>% filter(class == 1)
# 2반인 경우만 추출
exam %>% filter(class == 2)
# 1반이 아닌 경우
exam %>% filter(class != 1)
# 1반 이면서 수학 점수가 50점 이상인 경우
exam %>% filter(class == 1 & math >= 50)
# 수학 점수가 90점 이상이거나 영어점수가 90점 이상인 경우
exam %>% filter(math >= 90 | english >= 90)
#filter처리한 것임으로 원본데이터 유지
exam
```

실습 - 저장

```
# 1, 3, 5 반에 해당되면 추출
exam %>% filter(class == 1 | class == 3 | class == 5)
# 1, 3, 5 반에 해당하면 추출
exam %>% filter(class %in% c(1,3,5))

# class가 1인 행 추출, class1에 할당
class1 <- exam %>% filter(class == 1)
# class가 2인 행 추출, class2에 할당
class2 <- exam %>% filter(class == 2)
# 1반 수학 점수 평균 구하기
mean(class1$math)
# 2반 수학 점수 평균 구하기
mean(class2$math)
```

연산자

논리 연산자	기능
<	작다
<=	작거나 같다
>	크다
>=	크거나 같다
==	같다
!=	같지 않다
|	또는
&	그리고
%in%	매칭 확인

산술 연산자	기능
+	더하기
-	빼기
*	곱하기
/	나누기
^ , **	제곱
%/%	나눗셈의 몫
%%	나눗셈의 나머지

Quiz

❖ mpg 데이터를 이용해 분석 문제를 해결해 보세요.

- Q1. 자동차 배기량에 따라 고속도로 연비가 다른지 알아보려고 합니다. displ(배기량)이 4 이하인 자동차와 5 이상인 자동차 중 어떤 자동차의 hwy(고속도로 연비)가 평균적으로 더 높은지 알아보세요.
- Q2. 자동차 제조 회사에 따라 도시 연비가 다른지 알아보려고 합니다. "audi"와 "toyota" 중 어느 manufacturer(자동차 제조 회사)의 cty(도시 연비)가 평균적으로 더 높은지 알아보세요.
- Q3. "chevrolet", "ford", "honda" 자동차의 고속도로 연비 평균을 알아보려고 합니다. 이 회사들의 자동차를 추출한 뒤 hwy 전체 평균을 구해보세요

select

특정 행 추출

id	class	english	science
1	2	98	50
2	1	97	60
3	2	86	78
4	1	98	58
5	1	80	65
6	2	89	98

class	english
2	98
1	97
2	86
1	98
1	80
2	89

실습

```
# math 추출
exam %>% select(math)
# english 추출
exam %>% select(english)
# class, math, english 변수 추출
exam %>% select(class, math, english)
# math 제외
exam %>% select(-math)
# math, english 제외
exam %>% select(-math, -english)
```

실습

```
# class가 1인 행만 추출한 다음 english 추출
exam %>% filter(class == 1) %>% select(english)
#가독성 높이기
exam %>%
  filter(class == 1) %>%  # class가 1인 행 추출
  select(english)         # english 추출
#일부만 추출
exam %>%
  select(id, math) %>%  # id, math 추출
  head                  # 앞부분 6행까지 추출
exam %>%
  select(id, math) %>%  # id, math 추출
  head(10)              # 앞부분 10행까지 추출
```

Quiz

❖ mpg 데이터를 이용해서 분석 문제를 해결해보세요.

- Q1. mpg 데이터는 11개 변수로 구성되어 있습니다. 이 중 일부만 추출해서 분석에 활용하려고 합니다. mpg 데이터에서 class(자동차 종류), cty(도시 연비) 변수를 추출해 새로운 데이터를 만드세요. 새로 만든 데이터의 일부를 출력해서 두 변수로만 구성되어 있는지 확인하세요.
- Q2. 자동차 종류에 따라 도시 연비가 다른지 알아보려고 합니다. 앞에서 추출한 데이터를 이용해서 class(자동차 종류)가 "suv"인 자동차와 "compact"인 자동차 중 어떤 자동차의 cty(도시 연비)가 더 높은지 알아보세요.

arrange

정렬하기

id	english	science
1	98	50
2	97	60
3	86	78
4	98	58
5	80	65
6	89	98

id	english	science
6	89	98
5	86	78
4	80	65
3	97	60
2	98	58
1	98	50

실습

```
# math 오름차순 정렬
exam %>% arrange(math)
# math 내림차순 정렬
exam %>% arrange(desc(math))
# class 및 math 오름차순 정렬
exam %>% arrange(class, math)
```

Quiz

❖ mpg 데이터를 이용해서 분석 문제를 해결해보세요.

- "audi"에서 생산한 자동차 중에 어떤 자동차 모델의 hwy(고속도로 연비)가 높은지 알아보려고 합니다. "audi"에서 생산한 자동차 중 hwy가 1~5위에 해당하는 자동차의 데이터를 출력하세요.

mutate

파생 변수

id	english	science
1	98	50
2	97	60
3	86	78
4	98	58
5	80	65
6	89	98

id	english	science	total
1	98	50	148
2	97	60	157
3	86	78	164
4	98	58	156
5	80	65	145
6	89	98	187

실습

```
exam %>%
  mutate(total = math + english + science) %>%      # 총합 변수 추가
  head                                              # 일부 추출
exam %>%
  mutate(total = math + english + science,          # 총합 변수 추가
         mean = (math + english + science)/3) %>%   # 총평균 변수 추가
  head                                              # 일부 추출
exam %>%
  mutate(test = ifelse(science >= 60, "pass", "fail")) %>%     #조건처리
  head                                                         #일부 출력
exam %>%
  mutate(total = math + english + science) %>%      # 총합 변수 추가
  arrange(total) %>%                                # 총합 변수 기준 정렬
  head                                              # 일부 추출
```

Quiz

- mpg 데이터를 이용해서 분석 문제를 해결해보세요.
- mpg 데이터는 연비를 나타내는 변수가 hwy(고속도로 연비), cty(도시 연비) 두 종류로 분리되어 있습니다. 두 변수를 각각 활용하는 대신 하나의 통합 연비 변수를 만들어 분석하려고 합니다.
  - Q1. mpg 데이터 복사본을 만들고, cty와 hwy를 더한 '합산 연비 변수'를 추가하세요.
  - Q2. 앞에서 만든 '합산 연비 변수'를 2로 나눠 '평균 연비 변수'를 추가세요.
  - Q3. '평균 연비 변수'가 가장 높은 자동차 3종의 데이터를 출력하세요.
  - Q4. 1~3번 문제를 해결할 수 있는 하나로 연결된 dplyr 구문을 만들어 출력하세요. 데이터는 복사본 대신 mpg 원본을 이용하세요.

summarise

집단별 요약하기

class	english	science
2	98	50
1	97	60
2	86	78
1	98	58
1	80	65
2	89	98

class	english	science
1	97	60
1	98	58
1	80	65

mean(science)	
class 1	61.0

class	english	science
2	98	50
2	86	78
2	89	98

mean(science)	
class 2	75.3

mean(science)	
class 1	61.0
class 2	75.3

실습

```
# math 평균 산출
exam %>% summarise(mean_math = mean(math))


exam %>%
  group_by(class) %>%                  # class별로 분리
  summarise(mean_math = mean(math))  # math 평균 산출


exam %>%
  group_by(class) %>%                    # class별로 분리
  summarise(mean_math = mean(math),      # math 평균
            sum_math = sum(math),        # math 합계
            median_math = median(math), # math 중앙값
            n = n())                     # 학생 수
mpg %>%
  group_by(manufacturer, drv) %>%      # 회사별, 구방방식별 분리
  summarise(mean_cty = mean(cty)) %>%  # cty 평균 산출
  head(10)                             # 일부 출력
```

Quiz

❖ 문제) 회사별로 "suv" 자동차의 도시 및 고속도로 통합 연비 평균을 구해 내림차순으로 정렬하고, 1~5위까지 출력하기

절차	기능	dplyr 함수
1	회사별로 분리	group_by()
2	suv 추출	filter()
3	통합 연비 변수 생성	mutate()
4	통합 연비 평균 산출	summarise()
5	내림차순 정렬	arrange()
6	1~5위까지 출력	head()

Quiz

❖ mpg 데이터를 이용해서 분석 문제를 해결해 보세요.

- Q1. mpg 데이터의 class는 "suv", "compact" 등 자동차를 특징에 따라 일곱 종류로 분류한 변수입니다. 어떤 차종의 연비가 높은지 비교해보려고 합니다. class별 cty 평균을 구해보세요.
- Q2. 앞 문제의 출력 결과는 class 값 알파벳 순으로 정렬되어 있습니다. 어떤 차종의 도시 연비가 높은지 쉽게 알아볼 수 있도록 cty 평균이 높은 순으로 정렬해 출력하세요.
- Q3. 어떤 회사 자동차의 hwy(고속도로 연비)가 가장 높은지 알아보려고 합니다. hwy 평균이 가장 높은 회사 세 곳을 출력하세요.
- Q4. 어떤 회사에서 "compact"(경차) 차종을 가장 많이 생산하는지 알아보려고 합니다. 각 회사별 "compact" 차종 수를 내림차순으로 정렬해 출력하세요

데이터 합치기

left_join, bind_rows

id	midterm
1	60
2	80
3	70

+

id	final
1	70
2	83
3	65

=

id	midterm	final
1	60	70
2	80	83
3	70	65

id	test
1	60
2	80
3	70

+

id	test
4	70
5	83
6	65

=

id	test
1	60
2	80
3	70
4	70
5	83
6	65

세로로 합치기

실습

```
# 중간고사 데이터 생성
test1 <- data.frame(id = c(1, 2, 3, 4, 5), midterm = c(60, 80, 70, 90, 85))
# 기말고사 데이터 생성
test2 <- data.frame(id = c(1, 2, 3, 4, 5), final = c(70, 83, 65, 95, 80))

# id 기준으로 합쳐 total에 할당
total <- left_join(test1, test2, by = "id")
# total 출력
total
```

실습

```
exam
#반별 담당선생님 배정
name <- data.frame(class = c(1, 2, 3, 4, 5),
                   teacher = c("kim", "lee", "park", "choi", "jung"))
#이전 데이터와 합치기
exam_new <- left_join(exam, name, by = "class")
exam_new
```

실습

```
# 학생 1~5번 시험 데이터 생성
group_a <- data.frame(id = c(1, 2, 3, 4, 5),
                      test = c(60, 80, 70, 90, 85))

# 학생 6~10번 시험 데이터 생성
group_b <- data.frame(id = c(6, 7, 8, 9, 10),
                      test = c(70, 83, 65, 95, 80))
group_all <- bind_rows(group_a, group_b)  # 데이터 합쳐서 group_all에 할당
group_all
```

Quiz

- mpg 데이터를 이용해서 분석 문제를 해결해 보세요.
- mpg 데이터의 fl 변수는 자동차에 사용하는 연료(fuel)를 의미합니다. 아래는 자동차 연료별 가격을 나타낸 표입니다.

fl	연료 종류	가격(갤런당 USD)
c	CNG	2.35
d	diesel	2.38
e	ethanol E85	2.11
p	premium	2.76
r	regular	2.22

  - 이 정보를 이용해서 연료(f1)와 가격(price_f1)으로 구성된 데이터 프레임을 만들어 보세요

Quiz

- Q1. mpg 데이터에는 연료 종류를 나타낸 fl 변수는 있지만 연료 가격을 나타낸 변수는 없습니다. 위에서 만든 fuel 데이터를 이용해서 mpg 데이터에 price_fl(연료 가격) 변수를 추가하세요.
- Q2. 연료 가격 변수가 잘 추가됐는지 확인하기 위해서 model, fl, price_fl 변수를 추출해 앞부분 5행을 출력해 보세요.

Quiz

❖ 미국 동북중부 437개 지역의 인구통계 정보를 담고 있는 midwest 데이터를 사용해 데이터 분석 문제를 해결해 보세요. midwest는 ggplot2 패키지에 들어 있습니다.

- 문제1. popadults는 해당 지역의 성인 인구, poptotal은 전체 인구를 나타냅니다. midwest 데이터에 '전체 인구 대비 미성년 인구 백분율' 변수를 추가하세요.
- 문제2. 미성년 인구 백분율이 가장 높은 상위 5개 county(지역)의 미성년 인구 백분율을 출력하세요.
- 문제3. 분류표의 기준에 따라 미성년 비율 등급 변수를 추가하고, 각 등급에 몇 개의 지역이 있는지 알아보세요.

분류	기준
large	40% 이상
middle	30% ~ 40% 미만
small	30% 미만

- 문제4. popasian은 해당 지역의 아시아인 인구를 나타냅니다. '전체 인구 대비 아시아인 인구 백분율' 변수를 추가하고, 하위 10개 지역의 state(주), county(지역명), 아시아인 인구 백분율을 출력하세요