05 dplyr를 이용한 데이터 처리

학습내용

- dplyr 패키지의 함수를 알아보기

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

In [78]:

```
library(dplyr)
titanic <- read.csv("D:WWdatasetWWtitanic_dataWWtr_mod.csv")
head(titanic,10)
```

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabi
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.2500	N
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	71.2833	C8
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 3101282	7.9250	N
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.1000	C12
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.0500	N
6	0	3	Moran, Mr. James	male	29	0	0	330877	8.4583	N
7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E4
8	0	3	Palsson, Master. Gosta Leonard	male	2	3	1	349909	21.0750	N
9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27	0	2	347742	11.1333	N
10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14	1	0	237736	30.0708	N

In [79]:

```
names(titanic)
```

'PassengerId' 'Survived' 'Pclass' 'Name' 'Sex' 'Age' 'SibSp' 'Parch' 'Ticket' 'Fare' 'Cabin' 'Embarked'

In [80]:

```
table(titanic$Pclass)
```

```
  1   2   3 
216 184 491 
```

titanic에서 Pclass가 1인 값만 가져오기

- filter(조건)
- titanic %>% filter(Pclass==1)

In [81]:

```
dat <- titanic %>% filter(Pclass==1)
head(dat,10)
```

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	71.2833	C85
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.1000	C123
7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46
12	1	1	Bonnell, Miss. Elizabeth	female	58	0	0	113783	26.5500	C103
24	1	1	Sloper, Mr. William Thompson	male	28	0	0	113788	35.5000	A6
28	0	1	Fortune, Mr. Charles Alexander	male	19	3	2	19950	263.0000	C23 C25 C27
31	0	1	Uruchurtu, Don. Manuel E	male	40	0	0	PC 17601	27.7208	NA
32	1	1	Spencer, Mrs. William Augustus (Marie Eugenie)	female	29	1	0	PC 17569	146.5208	B78
35	0	1	Meyer, Mr. Edgar Joseph	male	28	1	0	PC 17604	82.1708	NA
36	0	1	Holverson, Mr. Alexander Oskar	male	42	1	0	113789	52.0000	NA

(ex) 5-1 실습해 보기

- filter를 이용하여 Embarked가 C인 경우 출력해보기
- filter를 이용하여 Embarked가 C가 아닌 경우 행 보기

In [82]:

```
# 1등급에 탄 30세 미만인 사람들은 몇몇이나?
dat <- titanic %>% filter(Pclass==1 & Age <=30)   # 클래스가 1이면서 30점 미만 출력
count(dat)
dat
```

n
91

In [83]:

```
# 1등급에 항구별 사람들?
dat <- titanic %>% filter(Pclass==1 & Embarked=='S')   # 클래스가 1이면서 30점 미만 출력
count(dat)
dat <- titanic %>% filter(Pclass==1 & Embarked=='C')   # 클래스가 1이면서 30점 미만 출력
count(dat)
dat <- titanic %>% filter(Pclass==1 & Embarked=='Q')   # 클래스가 1이면서 30점 미만 출력
count(dat)
```

n
129

n
85

n
2

1,3,5반에 해당되는 친구 출력 %in% 이용

- filter (컬럼명 %in% c(1,3,5))

In [84]:

```
## 2등급, 3등급
dat <- titanic %>% filter( Pclass %in% c(2,3) )
count(dat)
dim(titanic)
is(dim(titanic))
dim(titanic)[1]   # 행 출력
```

n
675

891 12

'integer' 'numeric' 'vector' 'data.frameRowLabels'

891

In [85]:

```
count(dat) / dim(titanic)[1]
```

n
0.7575758

5-2 필요한 변수만 추출하기

- select(변수명)

In [86]:

```
names(titanic)
```

'PassengerId' 'Survived' 'Pclass' 'Name' 'Sex' 'Age' 'SibSp' 'Parch' 'Ticket' 'Fare'
'Cabin' 'Embarked'

In [87]:

```
dat <- titanic %>% select(Pclass, Name, Sex)
head(dat,10)
```

Pclass	Name	Sex
3	Braund, Mr. Owen Harris	male
1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female
3	Heikkinen, Miss. Laina	female
1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female
3	Allen, Mr. William Henry	male
3	Moran, Mr. James	male
1	McCarthy, Mr. Timothy J	male
3	Palsson, Master. Gosta Leonard	male
3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female
2	Nasser, Mrs. Nicholas (Adele Achem)	female

In [88]:

```
sel <- c('Survived', 'Pclass', 'Name', 'Sex', 'Age', 'SibSp')
dat <- titanic %>% select(sel)
head(dat,10)
```

Survived	Pclass	Name	Sex	Age	SibSp
0	3	Braund, Mr. Owen Harris	male	22	1
1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1
1	3	Heikkinen, Miss. Laina	female	26	0
1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1
0	3	Allen, Mr. William Henry	male	35	0
0	3	Moran, Mr. James	male	29	0
0	1	McCarthy, Mr. Timothy J	male	54	0
0	3	Palsson, Master. Gosta Leonard	male	2	3
1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27	0
1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14	1

변수 제외

In [90]:

```
dat <- titanic %>% select(-Survived)
head(dat,10)
```

PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embark
1	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.2500	NA	
2	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	71.2833	C85	
3	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 3101282	7.9250	NA	
4	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.1000	C123	
5	3	Allen, Mr. William Henry	male	35	0	0	373450	8.0500	NA	
6	3	Moran, Mr. James	male	29	0	0	330877	8.4583	NA	
7	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46	
8	3	Palsson, Master. Gosta Leonard	male	2	3	1	349909	21.0750	NA	
9	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27	0	2	347742	11.1333	NA	
10	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14	1	0	237736	30.0708	NA	

In [91]:

```
dat <- titanic %>% select(-PassengerId, -Survived)
head(dat,10)
```

Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.2500	NA	S
1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	71.2833	C85	C
3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 3101282	7.9250	NA	S
1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.1000	C123	S
3	Allen, Mr. William Henry	male	35	0	0	373450	8.0500	NA	S
3	Moran, Mr. James	male	29	0	0	330877	8.4583	NA	Q
1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46	S
3	Palsson, Master. Gosta Leonard	male	2	3	1	349909	21.0750	NA	S
3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27	0	2	347742	11.1333	NA	S
2	Nasser, Mrs. Nicholas (Adele Achem)	female	14	1	0	237736	30.0708	NA	C

filter()와 select() 합치기

PClass 1등급의 Survived 를 뺀 열 전체 보기

In [92]:

```
names(titanic)
```

'PassengerId' 'Survived' 'Pclass' 'Name' 'Sex' 'Age' 'SibSp' 'Parch' 'Ticket' 'Fare' 'Cabin' 'Embarked'

In [93]:

```
sel <- c('Survived')
dat <- titanic %>% filter(Pclass==1) %>% select(-sel)
head(dat,10)
```

PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarke
2	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	71.2833	C85	
4	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.1000	C123	
7	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46	
12	1	Bonnell, Miss. Elizabeth	female	58	0	0	113783	26.5500	C103	
24	1	Sloper, Mr. William Thompson	male	28	0	0	113788	35.5000	A6	
28	1	Fortune, Mr. Charles Alexander	male	19	3	2	19950	263.0000	C23 C25 C27	
31	1	Uruchurtu, Don. Manuel E	male	40	0	0	PC 17601	27.7208	NA	
32	1	Spencer, Mrs. William Augustus (Marie Eugenie)	female	29	1	0	PC 17569	146.5208	B78	
35	1	Meyer, Mr. Edgar Joseph	male	28	1	0	PC 17604	82.1708	NA	
36	1	Holverson, Mr. Alexander Oskar	male	42	1	0	113789	52.0000	NA	

PClass의 1등급 5행만 보기

In [95]:

```
titanic %>% filter(Pclass==1) %>% head(5)
```

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	71.2833	C85
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.1000	C123
7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46
12	1	1	Bonnell, Miss. Elizabeth	female	58	0	0	113783	26.5500	C103
24	1	1	Sloper, Mr. William Thompson	male	28	0	0	113788	35.5000	A6

더해보기 p138

5-3 나이 순서대로 정렬해 보기

- arrange(기준컬럼)

In [96]:

```
titanic %>% arrange(Age) %>% head(10)
```

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Ca
804	1	3	Thomas, Master. Assad Alexander	male	0.42	0	1	2625	8.5167	
756	1	2	Hamalainen, Master. Viljo	male	0.67	1	1	250649	14.5000	
470	1	3	Baclini, Miss. Helene Barbara	female	0.75	2	1	2666	19.2583	
645	1	3	Baclini, Miss. Eugenie	female	0.75	2	1	2666	19.2583	
79	1	2	Caldwell, Master. Alden Gates	male	0.83	0	2	248738	29.0000	
832	1	2	Richards, Master. George Sibley	male	0.83	1	1	29106	18.7500	
306	1	1	Allison, Master. Hudson Trevor	male	0.92	1	2	113781	151.5500	C C
165	0	3	Panula, Master. Eino Viljami	male	1.00	4	1	3101295	39.6875	
173	1	3	Johnson, Miss. Eleanor Ileen	female	1.00	1	1	347742	11.1333	
184	1	2	Becker, Master. Richard F	male	1.00	2	1	230136	39.0000	

In [97]:

```
titanic %>% arrange(Age) %>% head
```

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabi
804	1	3	Thomas, Master. Assad Alexander	male	0.42	0	1	2625	8.5167	N/
756	1	2	Hamalainen, Master. Viljo	male	0.67	1	1	250649	14.5000	N/
470	1	3	Baclini, Miss. Helene Barbara	female	0.75	2	1	2666	19.2583	N/
645	1	3	Baclini, Miss. Eugenie	female	0.75	2	1	2666	19.2583	N/
79	1	2	Caldwell, Master. Alden Gates	male	0.83	0	2	248738	29.0000	N/
832	1	2	Richards, Master. George Sibley	male	0.83	1	1	29106	18.7500	N/

In [98]:

```
### 합계를 구해서 이를 이용하여 정렬하기
titanic$SibSp_Parch <- titanic$SibSp + titanic$Parch
head(titanic, 10)
```

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabi
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.2500	N
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	71.2833	C8
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 3101282	7.9250	N
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.1000	C12
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.0500	N
6	0	3	Moran, Mr. James	male	29	0	0	330877	8.4583	N
7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E4
8	0	3	Palsson, Master. Gosta Leonard	male	2	3	1	349909	21.0750	N
9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27	0	2	347742	11.1333	N
10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14	1	0	237736	30.0708	N

In [99]:

```
titanic %>% arrange(desc(SibSp_Parch)) %>% head(10) # 내림 차순 정렬
```

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Er
160	0	3	Sage, Master. Thomas Henry	male	29	8	2	CA. 2343	69.55	NA	
181	0	3	Sage, Miss. Constance Gladys	female	29	8	2	CA. 2343	69.55	NA	
202	0	3	Sage, Mr. Frederick	male	29	8	2	CA. 2343	69.55	NA	
325	0	3	Sage, Mr. George John Jr	male	29	8	2	CA. 2343	69.55	NA	
793	0	3	Sage, Miss. Stella Anna	female	29	8	2	CA. 2343	69.55	NA	
847	0	3	Sage, Mr. Douglas Bullen	male	29	8	2	CA. 2343	69.55	NA	
864	0	3	Sage, Miss. Dorothy Edith "Dolly"	female	29	8	2	CA. 2343	69.55	NA	
60	0	3	Goodwin, Master. William Frederick	male	11	5	2	CA 2144	46.90	NA	
72	0	3	Goodwin, Miss. Lillian Amy	female	16	5	2	CA 2144	46.90	NA	
387	0	3	Goodwin, Master. Sidney Leonard	male	1	5	2	CA 2144	46.90	NA	

실습과제 5-3

- titanic 데이터를 이용해서 분석 문제를 해결해 보시오.
- 어떤 경우에 생존자가 가장 많은가? (여성/남성, 나이대, 형제나자매수, Embarked 등)

5-4 파생변수 추가하기

- mutate(변수명)
- titanic$SibSp_Parch_m 생성
- titanic %>% mutate(SibSp_Parch_m= SibSp + Parch) %>% head

In [100]:

```
titanic %>% mutate(SibSp_Parch_m = SibSp + Parch) %>% head
```

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabi
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.2500	N
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	71.2833	C8
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 3101282	7.9250	N
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.1000	C12
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.0500	N
6	0	3	Moran, Mr. James	male	29	0	0	330877	8.4583	N

여러개의 변수를 생성하기

In [101]:

```
titanic %>% mutate(EM_C = ifelse(Embarked=='C',1,0), EM_S = ifelse(Embarked=='S',1,0)) %>% head
```

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabi
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.2500	N
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	71.2833	C8
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 3101282	7.9250	N
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.1000	C12
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.0500	N
6	0	3	Moran, Mr. James	male	29	0	0	330877	8.4583	N

In [102]:

```
summary(titanic$Age)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.42   22.00   29.00   29.56   35.00   80.00
```

5-5 집단별로 요약하기

- group_by()
- summarise()

exam의 math, english, science의 평균을 구해보자

- summarise(변수명 = mean(변수명))

In [103]:

```
titanic %>% summarise(mean_age = mean(Age))
```

mean_age
29.56024

In [104]:

```
titanic %>% summarise(mean_age = mean(Age),
                      mean_Fare = mean(Fare))
```

mean_age	mean_Fare
29.56024	32.20421

집단별로 분리해서 위의 내용을 출력해 보자.

- 그룹으로 묶고 이에 대한 요약값을 구하기
- 데이터 셋 %>% group_by(변수명) %>% summarise(...)

In [105]:

```
titanic %>% group_by(Pclass) %>% summarise(mean_age = mean(Age),
                      mean_Fare = mean(Fare))
```

Pclass	mean_age	mean_Fare
1	36.95102	84.15469
2	29.82516	20.66218
3	26.20961	13.67555

In [107]:

```
titanic %>% group_by(Pclass) %>%
          summarise(mean_age = mean(Age),
                    mean_Fare = mean(Fare),
                    mean_SibSp_Parch = mean(SibSp_Parch),
                    n=n())
```

Pclass	mean_age	mean_Fare	mean_SibSp_Parch	n
1	36.95102	84.15469	0.7731481	216
2	29.82516	20.66218	0.7826087	184
3	26.20961	13.67555	1.0081466	491

함수	설명
mean()	평균
sd()	표준편차
sum()	합계
median()	중앙값
min()	최솟값
max()	최댓값
n()	빈도

mpg 데이터 셋 불러오기

In [108]:

```
library(ggplot2)
```

In [109]:

```
mpg <- ggplot2::mpg
```

In [110]:

```
head(mpg)
```

manufacturer	model	displ	year	cyl	trans	drv	cty	hwy	fl	class
audi	a4	1.8	1999	4	auto(l5)	f	18	29	p	compact
audi	a4	1.8	1999	4	manual(m5)	f	21	29	p	compact
audi	a4	2.0	2008	4	manual(m6)	f	20	31	p	compact
audi	a4	2.0	2008	4	auto(av)	f	21	30	p	compact
audi	a4	2.8	1999	6	auto(l5)	f	16	26	p	compact
audi	a4	2.8	1999	6	manual(m5)	f	18	26	p	compact

제조사별 구동방식(drv)별 cty(도시연비)의 평균 산출하기

- drv(구동방식)
  - 4 사륜 구동
  - f 전륜 구동
  - r 후륜 구동

In [111]:

```
mpg %>% group_by(manufacturer, drv) %>%
    summarise(mean_cty = mean(cty)) %>%
    head(10)
```

manufacturer	drv	mean_cty
audi	4	16.81818
audi	f	18.85714
chevrolet	4	12.50000
chevrolet	f	18.80000
chevrolet	r	14.10000
dodge	4	12.00000
dodge	f	15.81818
ford	4	13.30769
ford	r	14.75000
honda	f	24.44444

실습과제 5-4

- 회사별로 분리한다.(group_by)
- class에서 suv 추출한다.(행) - filter()
- total(통합 연비) 파생변수 생성 - mutate()
- total(통합 연비) 의 평균 mean_total를 산출 - summarise()
- arrange() 내림 차순의 정렬
- 1~5위까지 : head()

In [112]:

```
mpg %>%
  group_by(manufacturer) %>%
  filter(class=="suv") %>%
  mutate(tot = (cty + hwy)/2) %>%
  summarise(mean_tot = mean(tot)) %>%
  arrange(desc(mean_tot)) %>% head(5)
```

manufacturer	mean_tot
subaru	21.91667
toyota	16.31250
nissan	15.87500
mercury	15.62500
jeep	15.56250

더해보기 p150

5-6 데이터 합치기

- dplyr:left_join()
- bind_rows()

In [113]:

```
kor <- data.frame(id=c(1,2,3,4,5),
                  kor=c(80,80,70,90,100))
math <- data.frame(id=c(1,2,3,4,5),
                   math=c(100,80,90,70,60))
eng <- data.frame(id=c(1,2,3,4,6),
                  eng=c(100,70,50,60,90))
```

- left_join(데이터셋1, 데이터셋2, by='컬럼명')

In [114]:

```
total <- left_join(kor, math, by='id')
total
```

id	kor	math
1	80	100
2	80	80
3	70	90
4	90	70
5	100	60

In [115]:

```
total_all <- left_join(total, eng, by='id')
total_all
```

id	kor	math	eng
1	80	100	100
2	80	80	70
3	70	90	50
4	90	70	60
5	100	60	NA

In [116]:

```
total_all2 <- left_join(eng, total, by='id')
total_all2
```

id	eng	kor	math
1	100	80	100
2	70	80	80
3	50	70	90
4	60	90	70
6	90	NA	NA

세로(행) 합치기, 행 추가

In [117]:

```
kor1 <- data.frame(id=c(1,2,3,4,5),
                   kor=c(80,80,70,90,100))
kor2 <- data.frame(id=c(6,7,8,9,10),
                   kor=c(80,80,70,90,100))
```

In [118]:

```
kor_all <- bind_rows(kor1, kor2)
```

In [119]:

```
kor_all
```

id	kor
1	80
2	80
3	70
4	90
5	100
6	80
7	80
8	70
9	90
10	100

변수가 kor로 동일하여 가능, 하지만 다를 경우는 어떻게 해야 하나? rename()를 이용하여 동일하게 한 이후에 합치기 가능