4. dplyr패키지

데이터 정제작업과 조작을 위한 dplyr패키지 주요함수 설명

I. dplyr패키지 개요

- Hadley Wickham이 작성한 패키지
- 비정형 데이터의 요약과 정제작업을 쉽게 빠르게 수행할 수 있음
- 데이터 조작을 수행하기 위한 다양한 함수들을 제공
- 벡터보다는 데이터 프레임 처리에 적합한 함수들로 구성

+ 참고) SQL과 비교

- SQL – 데이터 쿼리와 관리를 위한 언어
- dplyr – 데이터분석을 위해 디자인된 패키지 – Mysql, PostgreSQL, SQLite, BigQuery 를 지원

dplyr함수	유사한 SQL명령	기능
select( )	Select	컬럼 선택
filter()	where	행을 필터링(서브셋)
group_by()	Group by	데이터 그룹화
summarise() / aggregate()		데이터 요약 / 집계
arrange()	Order by	데이터 정렬
join()	JOIN	데이터프레임 조인
mutate()	Column alias	새로운 변수 생성

❑ dplyr 패키지 설치 및 기본함수 익히기

```
> install.packages("dplyr")
--- 현재 세션에서 사용할 CRAN 미러를 선택해 주세요 ---
'bindr', 'assertthat', 'bindrcpp', 'glue', 'pkgconfig', 'BH', '$
```

I. distinct() → 데이터 셋에서 중복된 행을 제거

- distinct(데이터셋) → 데이터셋의 모든 항목에 대해 중복체크
- distinct(데이터셋, 중복체크열, 옵션) →특정 열에 대한 중복체크

중복된 행을 제거

다른 열은 영향받지 않음

```
> df<-read.csv("d:\\mining-r\\exdata.csv",header=TRUE)
> result2<-distinct(df,Index,.keep_all=TRUE)
> result2
```

Index열에 대해 중복된 데이터 찾아서 제거

	Index	State	Y2002	Y2003	Y2004	Y2005	Y2006	Y2007
1	A	Alabama	1296530	1317711	1118631	1492583	1107408	1440134
2	C	California	1685349	1675807	1889570	1480280	1735069	1812546
3	D	Delaware	1330403	1268673	1706751	1403759	1441351	1300836
4	F	Florida	1964626	1468852	1419738	1362787	1339608	1278550
5	G	Georgia	1929009	1541565	1810773	1779091	1326846	1223770
6	H	Hawaii	1461570	1200280	1213993	1245931	1459383	1430465

2. select() – 특정 열을 추출

•select(데이터셋,추출할 열) → 해당 데이터 셋에서 추출할 열을 기술

```
> result3<-select(df,State,Y2013:Y2015)
> result3
                  State   Y2013   Y2014   Y2015
1               Alabama 1852841 1558906 1916661
2                Alaska 1985302 1580394 1979143
3               Arizona 1363279 1525866 1647724
4              Arkansas 1591896 1360959 1329341
5            California 1156536 1388461 1644607
6              Colorado 1178355 1383978 1330736
```

df에서 State, y2013~y2015까지 열을 추출

3. filter() – 사용자가 원하는 조건에 맞는 데이터 서브셋 구성함수

- 기본형식- filter(데이터셋, 조건)
- 데이터프레임 → 값이 를 만족하는 행만 모두 추출

```
> f1<-filter(df,Index %in% c('A','D'))
> f1
  Index                State   Y2002   Y2003   Y2004   Y2005   Y2006
1     A              Alabama 1296530 1317711 1118631 1492583 1107408
2     A               Alaska 1170302 1960378 1818085 1447852 1861639
3     A              Arizona 1742027 1968140 1377583 1782199 1102568
4     A             Arkansas 1485531 1994927 1119299 1947979 1669191
5     D             Delaware 1330403 1268673 1706751 1403759 1441351
6     D District of Columbia 1111437 1993741 1374643 1827949 1803852
```

df에서 Index열의 값이 'A','D'를 만족하는 모든 행을 추출

```
> f2<-filter(df,Index %in% c('A','D') & Y2015<1640000)
> f2
  Index                State   Y2002   Y2003   Y2004   Y2005   Y2006   Y2007
1     A             Arkansas 1485531 1994927 1119299 1947979 1669191 1801213
2     D             Delaware 1330403 1268673 1706751 1403759 1441351 1300836
3     D District of Columbia 1111437 1993741 1374643 1827949 1803852 1595981
    Y2008   Y2009   Y2010   Y2011   Y2012   Y2013   Y2014   Y2015
1 1188104 1628980 1669295 1928238 1216675 1591896 1360959 1329341
2 1762096 1553585 1370984 1318669 1984027 1671279 1803169 1627508
3 1193245 1739748 1707823 1353449 1979708 1912654 1782169 1410183
```

df에서 Index열의 값이 'A','D'를 만족하고 and,Y2015 값이 1640000미만을 만족하는 모든 행을 추출

```
> f4<-filter(df,grepl("Ne",State))
> f4
  Index          State   Y2002   Y2003   Y2004   Y2005   Y2006   Y2007
1     N       Nebraska 1885081 1309769 1425527 1240465 1500594 1278272
2     N         Nevada 1426117 1114500 1119707 1758830 1694526 1765826
3     N  New Hampshire 1419776 1854370 1195119 1990062 1645430 1286967
4     N     New Jersey 1605532 1141514 1613550 1181452 1541327 1156804
5     N     New Mexico 1819239 1226057 1935991 1124400 1723493 1475985
6     N       New York 1395149 1611371 1170675 1446810 1426941 1463171
```

grepl() -->grepl(패턴,텍스트) → 텍스트에서 특정 패턴을 찾아냄.

4. summarise() 함수 – 데이터 셋으로부터 특정 컬럼을 요약집계하는 함수

- 기본형식- summarise(데이터셋, 집계열변수명=집계함수(대상열)
- 데이터프레임 ➔ 데이터를 합산하여 결과를 열에 표시

```
> summarise(df,평균치2002=mean(Y2002))
  평균치2002
1    1566034
>
```

복수의 열에 대해서 평균과 표준편차를 일괄적으로 계산

```
> summarise_at(df,vars(Y2011:Y2015),funs(mean,sd))
  Y2011_mean Y2012_mean Y2013_mean Y2014_mean Y2015_mean Y2011_sd Y2012_sd
1    1574968    1591135    1530078    1583360    1588297 265721.6 283767.5
  Y2013_sd Y2014_sd Y2015_sd
1 282729.9 260155.4 274380.7
```

5. arrange()함수 – 데이터 정렬함수

- 기본형식- arrange(데이터셋, 정렬할 대상열)
- 데이터프레임 ➔ 데이터를 오름차순 정렬
- 내림차순 정렬 사용 → 데이터 내림차순 정렬

Index열 중심으로 오름차순 정렬하고, 같은 경우 Y2002 값이 큰 순서부터 표시되도록 정렬

```
> arrange(df1,Index,desc(Y2002))
  Index          State   Y2002   Y2003   Y2004   Y2005   Y2006
1     A        Arizona 1742027 1968140 1377583 1782199 1102568
2     A       Arkansas 1485531 1994927 1119299 1947979 1669191
3     A        Alabama 1296530 1317711 1118631 1492583 1107408
4     A         Alaska 1170302 1960378 1818085 1447852 1861639
5     C     California 1685349 1675807 1889570 1480280 1735069
6     C    Connecticut 1610512 1232844 1181949 1518933 1841266
```

6. group_by()함수 – 특정 열을 중심으로 데이터를 그룹화 함.

- 기본형식- group_by(데이터셋, 그룹화할 대상열)
- 데이터프레임 ➔ 별로 그룹화하겠다는 의미

```
> a=df1 %>% group_by(Index) %>% summarise(avg=mean(Y2015,na.rm=TRUE))
> a
# A tibble: 10 x 2
    Index     avg
   <fctr>   <dbl>
 1      A 1718217
 2      C 1564472
 3      D 1518846
 4      F 1170389
 5      G 1725470
 6      H 1150882
 7      I 1612542
 8      K 1649439
 9      L 1403857
10      M 1794366
> 
```

%>% 파이프 연산자
df1 대상으로 Index별로 그룹화 하고,y2015 평균을 계산한 후에 결과 보여줌

7. mutate()함수 – 새로운 변수를 생성

- 기본형식—mutate(데이터셋, 추가할 변수명=식)
- 데이터프레임 ➔ 에서 값을 뺀 차이를 에 저장

y2015-y2014를 계산하여 differ변수 생성
differ값이 300000초과하는 데이터 행을 추출

```
> a1<-mutate(df1,differ=Y2015-Y2014)
> a2<-filter(a1,differ>300000)
> a2
  Index       State   Y2002   Y2003   Y2004   Y2005   Y2006   Y2007   Y2008
1     A     Alabama 1296530 1317711 1118631 1492583 1107408 1440134 1945229
2     A      Alaska 1170302 1960378 1818085 1447852 1861639 1465841 1551826
3     M       Maine 1582720 1678622 1208496 1912040 1438549 1330014 1295877
4     M   Minnesota 1729921 1675204 1903907 1561839 1985692 1148621 1328133
5     M Mississippi 1983285 1292558 1631325 1943311 1354579 1731643 1428291
    Y2009   Y2010   Y2011   Y2012   Y2013   Y2014   Y2015 differ
1 1944173 1237582 1440756 1186741 1852841 1558906 1916661 357755
2 1436541 1629616 1230866 1512804 1985302 1580394 1979143 398749
3 1969163 1627262 1706080 1437088 1318546 1116792 1529233 412441
4 1890633 1995304 1575533 1910216 1972021 1515366 1864553 349187
5 1568049 1383227 1629132 1988270 1907777 1649668 1991232 341564
>
```