그룹별 집계

#01. 작업준비

패키지 가져오기

데이터 가져오기

#02. groupby() 메서드 사용하기

1. 기본 사용 방법

불필요한 컬럼을 제외하는 경우

년도별 교통사고 사망자수만 집계해야 할 경우

2. 두 개 이상의 변수에 대한 처리

집계 기준이 두 개인 경우

두 개 이상의 집계함수를 사용할 경우

#03. 사용자 정의 함수를 통한 집계

함수 만들기

#04. 인덱스 지정 해제

그룹별 집계

판다스(Pandas)의 `groupby()` 기능은 데이터를 그룹별로 분할하여 독립된 그룹에 대하여 별도로 데이터를 처리(혹은 적용)하거나 그룹별 통계량을 확인하고자 할 때 유용한 함수

> SQL의 `group by` 구절과 동일한 기능

#01. 작업준비

패키지 가져오기

```
from pandas import read_excel
```

데이터 가져오기

```
df = read_excel("https://data.hossam.kr/C02/traffic_acc.xlsx")
df
```

	년도	월	발생건수	사망자수	부상자수
0	2005	1	15494	504	25413
1	2005	2	13244	431	21635
2	2005	3	16580	477	25550

그룹별 집계

#01. 작업준비

패키지 가져오기

데이터 가져오기

#02. groupby() 메서드 사용하기

1. 기본 사용 방법

불필요한 컬럼을 제외하는 경우

년도별 교통사고 사망자수만 집계해야 할 경우

2. 두 개 이상의 변수에 대한 처리

집계 기준이 두 개인 경우

두 개 이상의 집계함수를 사용할 경우

#03. 사용자 정의 함수를 통한 집계

함수 만들기

#04. 인덱스 지정 해제

	년도	월	발생건수	사망자수	부상자수
3	2005	4	17817	507	28131
4	2005	5	19085	571	29808
...	...	...	...	...	...
163	2018	8	18335	357	27749
164	2018	9	18371	348	27751
165	2018	10	19738	373	28836
166	2018	11	19029	298	28000
167	2018	12	18010	323	26463

168 rows × 5 columns

#02. `groupby()` 메서드 사용하기

1. 기본 사용 방법

지정된 컬럼을 기준으로 같은 값을 갖는 행을 그룹으로 묶고 지정되지 않은 컬럼에 대한 집계 함수를 명시해야 한다.

함수	내용
count	데이터의 개수
sum	합계

그룹별 집계

#01. 작업준비

패키지 가져오기

데이터 가져오기

#02. groupby() 메서드 사용하기

1. 기본 사용 방법

불필요한 컬럼을 제외하는 경우

년도별 교통사고 사망자수만 집계해야 할 경우

2. 두 개 이상의 변수에 대한 처리

집계 기준이 두 개인 경우

두 개 이상의 집계함수를 사용할 경우

#03. 사용자 정의 함수를 통한 집계

함수 만들기

#04. 인덱스 지정 해제

함수	내용
mean	평균
median	중앙값
var, std	분산, 표준편차
min, max	최소, 최대값
unique, nunique	고유값, 고유값 개수
prod	곱
first, last	첫째, 마지막값

아래 결과에서는 `월`도 합계를 구한다.

집계전 제외해야 하는 컬럼은 `drop()`을 사용해서 삭제하거나 집계에 포함되어야 하는 항목만 `filter()`로 추출한 후 `groupby()`를 적용하는 것이 바람직

```
df.groupby('년도').sum()
```

	월	발생건수	사망자수	부상자수
년도				
2005	78	214171	6376	342233
2006	78	213745	6327	340229
2007	78	211662	6166	335906
2008	78	215822	5870	338962

그룹별 집계

#01. 작업준비

패키지 가져오기

데이터 가져오기

#02. groupby() 메서드 사용하기

1. 기본 사용 방법

불필요한 컬럼을 제외하는 경우

년도별 교통사고 사망자수만 집계해야 할 경우

2. 두 개 이상의 변수에 대한 처리

집계 기준이 두 개인 경우

두 개 이상의 집계함수를 사용할 경우

#03. 사용자 정의 함수를 통한 집계

함수 만들기

#04. 인덱스 지정 해제

	월	발생건수	사망자수	부상자수
년도				
2009	78	231990	5838	361875
2010	78	226878	5505	352458
2011	78	221711	5229	341391
2012	78	223656	5392	344565
2013	78	215354	5092	328711
2014	78	223552	4762	337497
2015	78	232035	4621	350400
2016	78	220917	4292	331720
2017	78	216335	4185	322829
2018	78	217148	3781	323037

불필요한 컬럼을 제외하는 경우

```
df2 = df.drop('월', axis=1)
df2
```

	년도	발생건수	사망자수	부상자수
0	2005	15494	504	25413
1	2005	13244	431	21635

그룹별 집계

#01. 작업준비

패키지 가져오기

데이터 가져오기

#02. groupby() 메서드 사용하기

1. 기본 사용 방법

불필요한 컬럼을 제외하는 경우

년도별 교통사고 사망자수만 집계해야 할 경우

2. 두 개 이상의 변수에 대한 처리

집계 기준이 두 개인 경우

두 개 이상의 집계함수를 사용할 경우

#03. 사용자 정의 함수를 통한 집계

함수 만들기

#04. 인덱스 지정 해제

	년도	발생건수	사망자수	부상자수
2	2005	16580	477	25550
3	2005	17817	507	28131
4	2005	19085	571	29808
...	...	...	...	...
163	2018	18335	357	27749
164	2018	18371	348	27751
165	2018	19738	373	28836
166	2018	19029	298	28000
167	2018	18010	323	26463

168 rows × 4 columns

```
df2.groupby('년도').sum()
```

	발생건수	사망자수	부상자수
년도			
2005	214171	6376	342233
2006	213745	6327	340229
2007	211662	6166	335906
2008	215822	5870	338962

그룹별 집계

#01. 작업준비

패키지 가져오기

데이터 가져오기

#02. groupby() 메서드 사용하기

1. 기본 사용 방법

불필요한 컬럼을 제외하는 경우

년도별 교통사고 사망자수만 집계해야 할 경우

2. 두 개 이상의 변수에 대한 처리

집계 기준이 두 개인 경우

두 개 이상의 집계함수를 사용할 경우

#03. 사용자 정의 함수를 통한 집계

함수 만들기

#04. 인덱스 지정 해제

	발생건수	사망자수	부상자수
년도			
2009	231990	5838	361875
2010	226878	5505	352458
2011	221711	5229	341391
2012	223656	5392	344565
2013	215354	5092	328711
2014	223552	4762	337497
2015	232035	4621	350400
2016	220917	4292	331720
2017	216335	4185	322829
2018	217148	3781	323037

년도별 교통사고 사망자수만 집계해야 할 경우

```
df3 = df.filter(['년도', '사망자수'])
df3
```

	년도	사망자수
0	2005	504
1	2005	431

그룹별 집계

#01. 작업준비

패키지 가져오기

데이터 가져오기

#02. groupby() 메서드 사용하기

1. 기본 사용 방법

불필요한 컬럼을 제외하는 경우

년도별 교통사고 사망자수만 집계해야 할 경우

2. 두 개 이상의 변수에 대한 처리

집계 기준이 두 개인 경우

두 개 이상의 집계함수를 사용할 경우

#03. 사용자 정의 함수를 통한 집계

함수 만들기

#04. 인덱스 지정 해제

	년도	사망자수
2	2005	477
3	2005	507
4	2005	571
...	...	...
163	2018	357
164	2018	348
165	2018	373
166	2018	298
167	2018	323

168 rows × 2 columns

2. 두 개 이상의 변수에 대한 처리

집계 기준이 두 개인 경우

```
df.groupby(['년도', '월']).mean()
```

		발생건수	사망자수	부상자수
년도	**월**			
2005	**1**	15494.0	504.0	25413.0

그룹별 집계

#01. 작업준비

패키지 가져오기

데이터 가져오기

#02. groupby() 메서드 사용하기

1. 기본 사용 방법

불필요한 컬럼을 제외하는 경우

년도별 교통사고 사망자수만 집계해야 할 경우

2. 두 개 이상의 변수에 대한 처리

집계 기준이 두 개인 경우

두 개 이상의 집계함수를 사용할 경우

#03. 사용자 정의 함수를 통한 집계

함수 만들기

#04. 인덱스 지정 해제

		발생건수	사망자수	부상자수
년도	월			
	2	13244.0	431.0	21635.0
	3	16580.0	477.0	25550.0
	4	17817.0	507.0	28131.0
	5	19085.0	571.0	29808.0
...	...	...	...	...
2018	8	18335.0	357.0	27749.0
	9	18371.0	348.0	27751.0
	10	19738.0	373.0	28836.0
	11	19029.0	298.0	28000.0
	12	18010.0	323.0	26463.0

168 rows × 3 columns

두 개 이상의 집계함수를 사용할 경우

```
df.filter(['년도', '발생건수']).groupby('년도').agg(['sum', 'max', 'min',
```

그룹별 집계

#01. 작업준비

패키지 가져오기

데이터 가져오기

#02. groupby() 메서드 사용하기

1. 기본 사용 방법

불필요한 컬럼을 제외하는 경우

년도별 교통사고 사망자수만 집계해야 할 경우

2. 두 개 이상의 변수에 대한 처리

집계 기준이 두 개인 경우

두 개 이상의 집계함수를 사용할 경우

#03. 사용자 정의 함수를 통한 집계

함수 만들기

#04. 인덱스 지정 해제

	발생건수			
	sum	max	min	mean
년도				
2005	214171	19757	13244	17847.583333
2006	213745	19877	14270	17812.083333
2007	211662	19264	14696	17638.500000
2008	215822	19926	14176	17985.166667
2009	231990	21440	15502	19332.500000
2010	226878	21575	15803	18906.500000
2011	221711	20952	14208	18475.916667
2012	223656	19750	16656	18638.000000
2013	215354	19797	14187	17946.166667
2014	223552	20760	14061	18629.333333
2015	232035	21587	14939	19336.250000
2016	220917	19918	15664	18409.750000
2017	216335	19891	14832	18027.916667
2018	217148	19738	16208	18095.666667

#03. 사용자 정의 함수를 통한 집계

그룹별 집계

#01. 작업준비

패키지 가져오기

데이터 가져오기

#02. groupby() 메서드 사용하기

1. 기본 사용 방법

불필요한 컬럼을 제외하는 경우

년도별 교통사고 사망자수만 집계해야 할 경우

2. 두 개 이상의 변수에 대한 처리

집계 기준이 두 개인 경우

두 개 이상의 집계함수를 사용할 경우

#03. 사용자 정의 함수를 통한 집계

함수 만들기

#04. 인덱스 지정 해제

함수 만들기

```
def myFunction(x):
    # x는 시리즈 - 컬럼 하나가 통째로 전달됨
    return x.max() - x.min()
```

```
df.drop('월', axis=1).groupby('년도').agg(['max', 'min', myFunction])
```

	발생건수			사망자수			부상자수		
	max	min	myFunction	max	min	myFunction	max	min	m
년도									
2005	19757	13244	6513	639	431	208	31603	21635	9
2006	19877	14270	5607	701	373	328	31270	22903	8
2007	19264	14696	4568	582	446	136	30532	23717	6
2008	19926	14176	5750	574	423	151	30935	23282	7
2009	21440	15502	5938	592	405	187	33255	24429	8
2010	21575	15803	5772	619	395	224	33282	24968	8
2011	20952	14208	6744	520	338	182	32133	22493	9
2012	19750	16656	3094	533	393	140	30163	25998	4
2013	19797	14187	5610	499	335	164	29676	22255	7
2014	20760	14061	6699	476	325	151	31199	21501	9

그룹별 집계

#01. 작업준비

패키지 가져오기

데이터 가져오기

#02. groupby() 메서드 사용하기

1. 기본 사용 방법

불필요한 컬럼을 제외하는 경우

년도별 교통사고 사망자수만 집계해야 할 경우

2. 두 개 이상의 변수에 대한 처리

집계 기준이 두 개인 경우

두 개 이상의 집계함수를 사용할 경우

#03. 사용자 정의 함수를 통한 집계

함수 만들기

#04. 인덱스 지정 해제

	발생건수			사망자수			부상자수		
	max	min	myFunction	max	min	myFunction	max	min	m
년도									
2015	21587	14939	6648	470	296	174	32436	22999	9
2016	19918	15664	4254	440	292	148	29635	24021	5
2017	19891	14832	5059	420	280	140	29371	22323	7
2018	19738	16208	3530	373	266	107	28836	24630	4

#04. 인덱스 지정 해제

`groupby()` 메서드에서 사용한 필드는 결과 데이터프레임의 인덱스로 지정되는 것이 기본 형태

`groupby()` 메서드에 `as_index=False` 파라미터를 추가하여 인덱스 지정을 방지할 수 있다.

```
df.drop('월', axis=1).groupby('년도', as_index=False).sum()
```

	년도	발생건수	사망자수	부상자수
0	2005	214171	6376	342233
1	2006	213745	6327	340229
2	2007	211662	6166	335906
3	2008	215822	5870	338962

그룹별 집계

#01. 작업준비

패키지 가져오기

데이터 가져오기

#02. groupby() 메서드 사용하기

1. 기본 사용 방법

불필요한 컬럼을 제외하는 경우

년도별 교통사고 사망자수만 집계해야 할 경우

2. 두 개 이상의 변수에 대한 처리

집계 기준이 두 개인 경우

두 개 이상의 집계함수를 사용할 경우

#03. 사용자 정의 함수를 통한 집계

함수 만들기

#04. 인덱스 지정 해제

	년도	발생건수	사망자수	부상자수
4	2009	231990	5838	361875
5	2010	226878	5505	352458
6	2011	221711	5229	341391
7	2012	223656	5392	344565
8	2013	215354	5092	328711
9	2014	223552	4762	337497
10	2015	232035	4621	350400
11	2016	220917	4292	331720
12	2017	216335	4185	322829
13	2018	217148	3781	323037