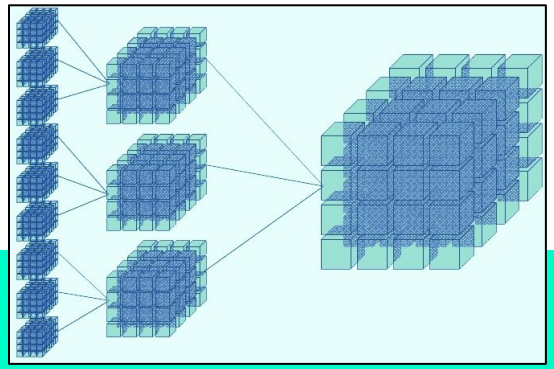
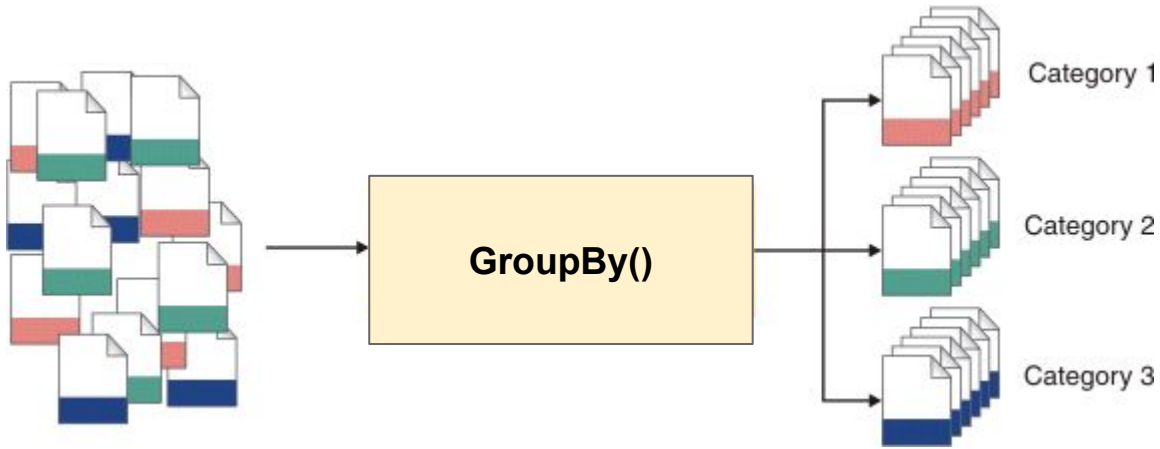
Data Aggregation and Group Operations.



Chapter 10.

Step 1: Categorize data

Popular groups include income, age, profession.



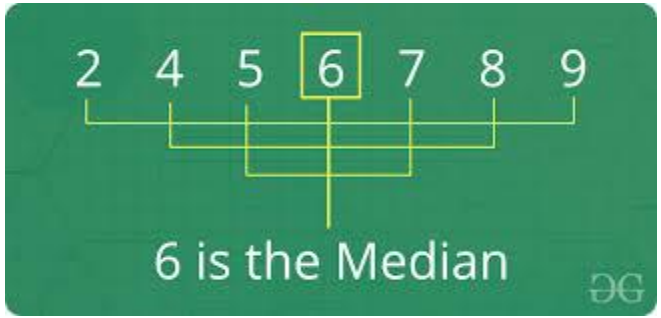
Step 2: Compute Group Summary Statistics

`count`	Number of non-NA values in the group
`sum`	Sum of non-NA values
`mean`	Mean of non-NA values
`median`	Arithmetic median of non-NA values
`std, var`	Unbiased (n - 1 denominator) standard deviation and variance
`min, max`	Minimum and maximum of non-NA values
`prod`	Product of non-NA values
`first, last`	First and last non-NA values

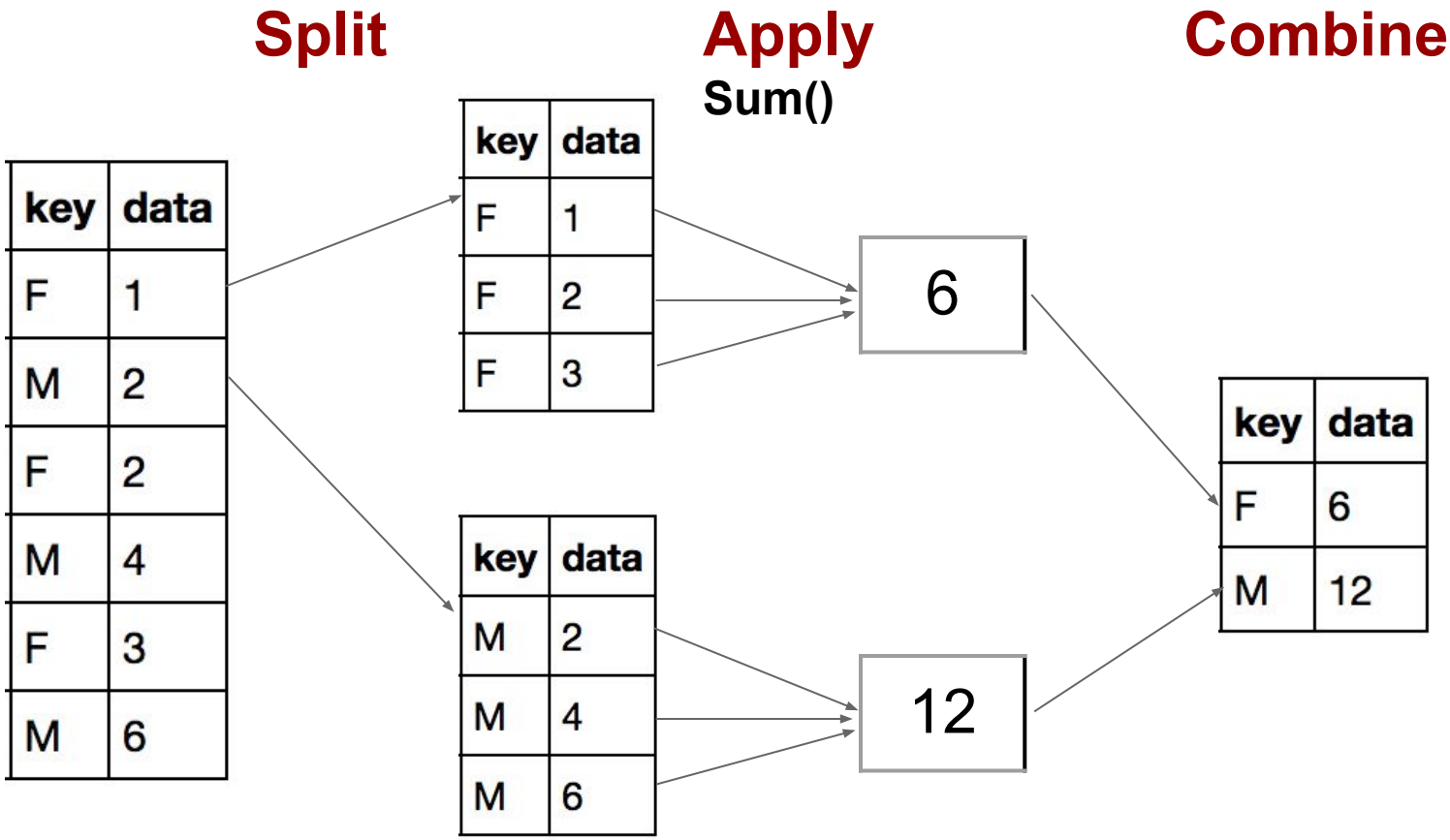
MEAN VS MEDIAN

The mean (average) of a data set is found by adding all numbers in the data set and then dividing by the number of values in the set.

The median is the middle value when a data set is ordered from least to greatest.



SPLIT-APPLY-COMBINE OF A SIMPLE AGGREGATION



Create a DataFrame

```
df = DataFrame({'key': ['F', 'M', 'F', 'M', 'F', 'M'],
                'data': [1, 2, 2, 4, 3, 6]},
               columns=['key', 'data'])
df
```

	key	data
0	F	1
1	M	2
2	F	2
3	M	4
4	F	3
5	M	6

Split - Apply - Combine

```
grouped = df['data'].groupby(df['key'])
grouped.sum()
```

```
key
F      6
M     12
Name: data, dtype: int64
```

Create a DataFrame II

```
df = DataFrame({'key1': ['F', 'F', 'F', 'M', 'F', 'M'],
                'key2': ['20-25', '20-25', '20-25', '25-30', '30-35', '30-35'
                'data1': [2, 2, 3, 3, 4, 4]},
                columns=['key1', 'key2', 'data1'])
df
```

	key1	key2	data1
0	F	20-25	2
1	F	20-25	2
2	F	20-25	3
3	M	25-30	3
4	F	30-35	4
5	M	30-35	4

If we group the data by using two keys...

the resulting Series has a **hierarchical index.**

```
grouped = df['data1'].groupby([df['key1'], df['key2']])
grouped.sum()
```

```
key1  key2
F     20-25    7
      30-35    4
M     25-30    3
      30-35    4
Name: data1, dtype: int64
```

Pivot Table

Month	(Multiple Items)				
Sum of Net Sales		Product			
Region	Salesman	FastCar	RapidZoo	SuperGlue	Grand Total
Middle	Joseph	$ 3,623	$ 4,782	$ 7,055	$ 15,460
	Lawrence	$ 5,908	$ 4,642	$ 4,593	$ 15,143
	Maria	$ 6,502	$ 3,969	$ 5,408	$ 15,879
	Matt	$ 4,170	$ 6,093	$ 5,039	$ 15,302
Middle Total		$ 20,203	$ 19,486	$ 22,095	$ 61,784
North	Joseph	$ 3,643	$ 5,846	$ 6,574	$ 16,063
	Lawrence	$ 4,456	$ 6,658	$ 7,685	$ 18,799
	Maria	$ 6,235	$ 4,616	$ 3,612	$ 14,463
	Matt	$ 3,868	$ 3,926	$ 3,254	$ 11,048
North Total		$ 18,202	$ 21,046	$ 21,125	$ 60,373
West	Joseph	$ 5,507	$ 5,186	$ 4,882	$ 15,575
	Lawrence	$ 4,082	$ 5,272	$ 6,124	$ 15,478
	Maria	$ 5,320	$ 5,461	$ 4,872	$ 15,653
	Matt	$ 6,737	$ 4,598	$ 4,233	$ 15,568
West Total		$ 21,646	$ 20,517	$ 20,111	$ 62,474
Grand Total		$ 60,251	$ 61,049	$ 63,331	$ 184,631

A pivot table is a data summarization tool.

It aggregates a table of data by one or more keys, arranging some of the group keys along the rows and some along the columns.

Pivot Table Example

sum() aggregate function

```
df.pivot_table('data1', index='key2', columns='key1', aggfunc='sum')
```

key1	F	M
key2		
20-25	7.0	NaN
25-30	NaN	3.0
30-35	4.0	4.0