C: Grouping and aggregating data

Computing statistics for groups within a dataset

Two conceptual steps:

1. Grouping
   Collect records into groups

   Example:
      Building **list of kW** used in each hour

   Purely organizational

2. Aggregating
   Compute group values from member data

   Example:
      Calculating **average kW** usage for each hour

   Applies a given calculation to each group
   Collapses the data: one value per group rather than per member

Implementation in Pandas:

1. Build grouped data using .groupby() method.

2. Select variables in grouped data and apply an aggregation function.

Example:

Step 0: Initial data

raw, index = id:

id	type	inc	age
1	A	50	32
2	B	80	40
3	B	30	20
4	A	70	27
5	B	88	50

Step 1: grouping

```
grouped = raw.groupby('type')
```

Internal groups:

type A:

id	inc	age
1	50	32
4	70	27

type B:

id	inc	age
2	80	40

3	30	20
5	88	50

Step 2: aggregating

Examples: applying functions to individual variables

```
mean_inc = grouped['inc'].mean()
```

type	inc
A	60
B	66

- Index will always be the aggregation variable(s)

```
med_age = grouped['age'].median()
```

type	age
A	29.5
B	40

What functions can be applied?

- A series method that returns a scalar

Examples:

.sum()	Sum
.quantile(0.25)	Value at the 25th percentile

- Several additional methods

  Examples:

.size()	Number of items in the group
.describe()	Descriptive statistics

Example: applying one function to all variables

```
means = grouped.mean()
```

Function applied to every column:

type	inc	age
A	60.0	29.500000
B	66.0	36.666667

Example: applying a function with multiple outputs to one variable

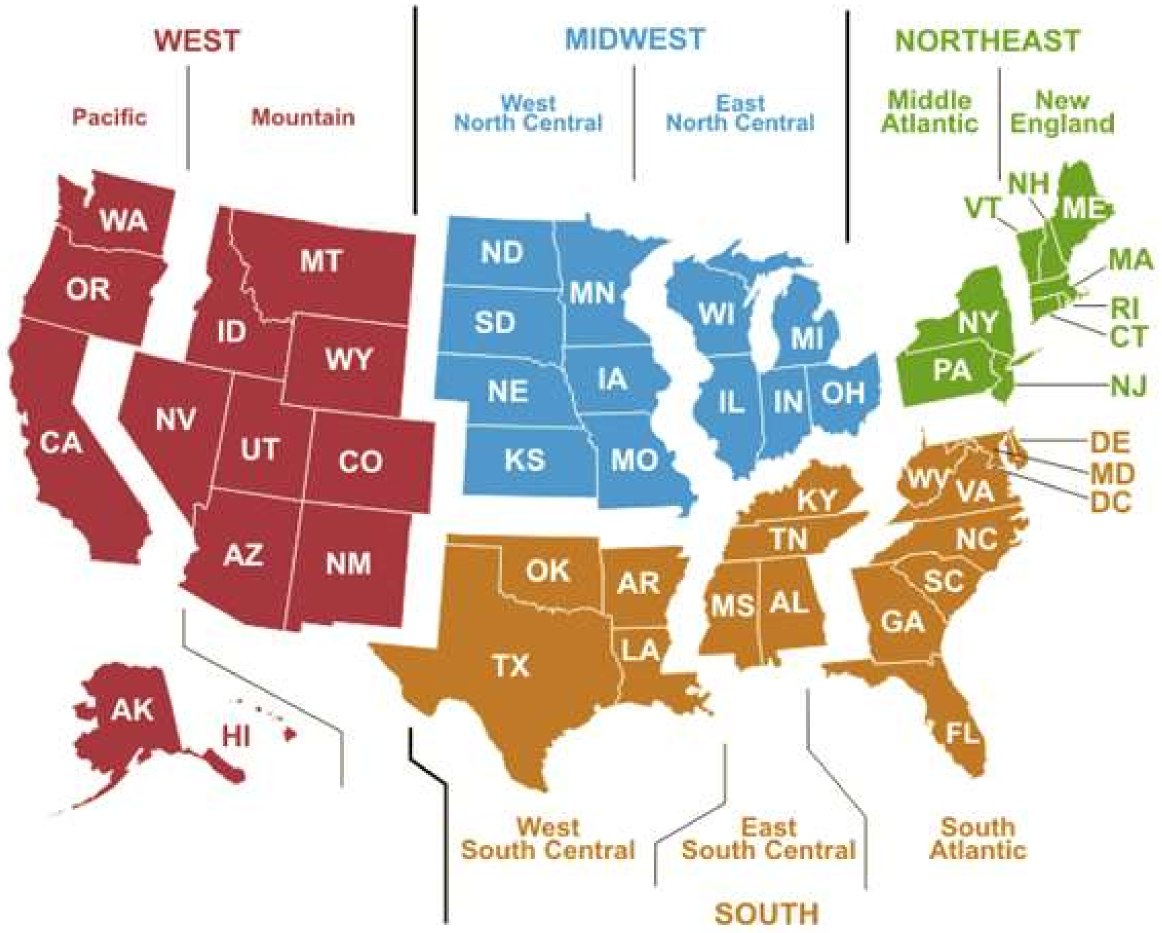
```
details = grouped['inc'].describe()
```

Generate multiple statistics per variable:

type	count	mean	std	min	25%	50%	75%	max
A	2.0	60.0	14.142136	50.0	55.0	60.0	65.0	70.0
B	3.0	66.0	31.432467	30.0	55.0	80.0	84.0	88.0

Background for demo.py

Census regions (4) and divisions (9):


WEST
Pacific
Mountain
MIDWEST
West North Central
East North Central
NORTHEAST
Middle Atlantic
New England
WA
OR
CA
MT
ID
WY
NV
UT
CO
AZ
NM
AK
HI
ND
SD
NE
KS
MN
IA
MO
WI
MI
IL
IN
OH
NY
PA
NJ
VT
NH
ME
MA
RI
CT
DE
MD
DC
WV
VA
NC
SC
GA
FL
KY
TN
MS
AL
OK
AR
TX
LA
West South Central
East South Central
South Atlantic
SOUTH