NYU Center for Data Science

DS-UA 112
Introduction to Data Science

Week 4: Lecture 2

Tables – Grouping and Pivoting Data

NYU Center for Data Science

What can we learn from rearranging a table to group related records?

DS-UA 112
Introduction to Data Science

Week 4: Lecture 2

Tables – Grouping and Pivoting Data

Adapted from Nolan, DeNero, Hug, Lau

Announcements

- Please check Week 4 agenda on NYU Classes
  - Homework 2
  - Lab 3
  - Survey 2
- Remember to post to Piazza

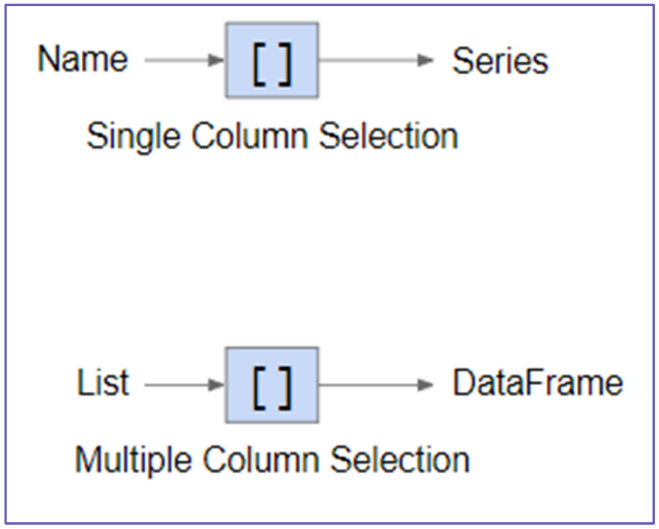
Remember to tag your answers on Gradescope!

Review

Access Columns

```
elections[["Candidate", "Party"]].head(6)
```

Year	Candidate	Party
1980	Reagan	Republican
1980	Carter	Democratic
1980	Anderson	Independent
1984	Reagan	Republican
1984	Mondale	Democratic
1988	Bush	Republican



- Use brackets with a label or list of labels for the columns

```
elections["Candidate"].head(6)
```

```
Year
1980       Reagan
1980       Carter
1980     Anderson
1984       Reagan
1984      Mondale
1988         Bush
Name: Candidate, dtype: object
```

Review

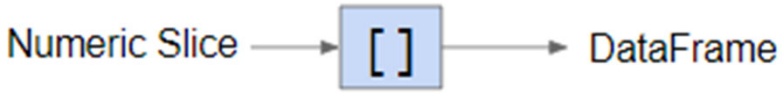
Access Rows

- Use brackets with numeric range or True/False

```
elections[elections['Party'] == 'Independent']
```

```
elections[[False, False, False, False, False,
           False, False, True, False, False,
           True, False, False, False, True,
           False, False, False, False, False,
           False, False, True]]
```

	Candidate	Party	%	Year	Result
7	Clinton	Democratic	43.0	1992	win
10	Clinton	Democratic	49.2	1996	win
14	Bush	Republican	47.9	2000	win
22	Trump	Republican	46.1	2016	win



(Multiple) Row Selection

```
elections[0:3]
```

Year	Candidate	Party	%	Result
1980	Reagan	Republican	50.7	win
1980	Carter	Democratic	41.0	loss
1980	Anderson	Independent	6.6	loss

Review

You must use & for “and”, | for “or”, ~ for “not”

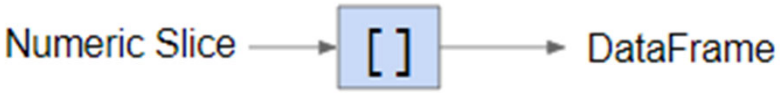
```
elections[(elections['Result'] == 'win')
          & (elections['%'] < 50)]
```

```
elections[[False, False, False, False, False,
           False, False, True, False, False,
           True, False, False, False, True,
           False, False, False, False, False,
           False, False, True]]
```

	Candidate	Party	%	Year	Result
7	Clinton	Democratic	43.0	1992	win
10	Clinton	Democratic	49.2	1996	win
14	Bush	Republican	47.9	2000	win
22	Trump	Republican	46.1	2016	win

Access Rows

- Use brackets with numeric range or True/False



(Multiple) Row Selection

```
elections[0:3]
```

Year	Candidate	Party	%	Result
1980	Reagan	Republican	50.7	win
1980	Carter	Democratic	41.0	loss
1980	Anderson	Independent	6.6	loss

Review

Access Rows and Columns

```
elections.loc[[0, 1, 2, 3, 4], ['Candidate','Party', 'Year']]
```

	Candidate	Party	Year
0	Reagan	Republican	1980
1	Carter	Democratic	1980
2	Anderson	Independent	1980
3	Reagan	Republican	1984
4	Mondale	Democratic	1984

- Use loc and iloc to specify both rows and columns
- loc accesses by
  - value of label
  - True or False

```
elections.iloc[0:3, 0:3]
```

- iloc accesses by
  - row number
  - column number

Review

Question

Link the following definitions to their corresponding Pandas data structure:

1. A sequence of row labels
2. Two-dimensional (tabular data)
3. One-dimensional (column data)

- ☐ Data Frame: 1, Series: 2, Index: 3
- ☐ Data Frame: 2, Series: 1, Index: 3
- ☐ Data Frame: 2, Series: 3, Index: 1
- ☐ Data Frame: 3, Series: 2, Index: 1

Review

Question

Link the following definitions to their corresponding Pandas data structure:

1. A sequence of row labels
2. Two-dimensional (tabular data)
3. One-dimensional (column data)

- ☐ Data Frame: 1, Series: 2, Index: 3
- ☐ Data Frame: 2, Series: 1, Index: 3
- ■ Data Frame: 2, Series: 3, Index: 1
- ☐ Data Frame: 3, Series: 2, Index: 1

Review

Question

Which of the following statements about Pandas Indices are false?

- ☐ Indices must integers
- ☐ Indices may be non-numeric, and are always unique
- ☐ Indices need not be unique, but must be numeric
- ☐ Indices need not be unique, and can be non-numeric

Review

Question

Which of the following statements about Pandas Indices are false?

- [x] Indices must integers
- [x] Indices may be non-numeric, and are always unique
- [x] Indices need not be unique, but must be numeric
- [ ] Indices need not be unique, and can be non-numeric

Review

Question

Which of the following statements regarding `iloc` are true?

- ☐ It is harder to make mistakes with iloc than with loc
- ☐ It is easier to read iloc code than loc code
- ☐ iloc doesn't use labels
- ☐ iloc is vulnerable to changes in the ordering of rows and columns in a Data Frame

Review

Question

Which of the following statements regarding `iloc` are true?

- ☐ It is harder to make mistakes with iloc than with loc
- ☐ It is easier to read iloc code than loc code
- ■ iloc doesn't use labels
- ■ iloc is vulnerable to changes in the ordering of rows and columns in a Data Frame

Review

Question

Which of the following statements regarding `iloc` are true?

- ☐ It is harder to make mistakes with iloc than with loc
- ☐ It is easier to read iloc code than loc code
- ■ iloc doesn't use labels
- ■ iloc is vulnerable to changes in the ordering of rows and columns in a Data Frame

Agenda

- Operations on Tables
  - Inspecting
  - Sorting
  - Summarizing
- Grouping and Pivoting
  - Find the most popular name in New York
  - Find all names that start with E.
  - Sort names by occurrence of dr and ea
  - Find the name whose popularity has changed the most.
  - Count the number of female and male babies born in each year

References

- Nolan, Lau, Gonzalez (Chapter 3.3)
- McKinney (Chapter 5, 8)

Grouping

How can we change the granularity of data in a table?

Key	Data
A	3
B	1
C	4
A	1
B	5
C	9
A	2
B	6
C	5

Grouping

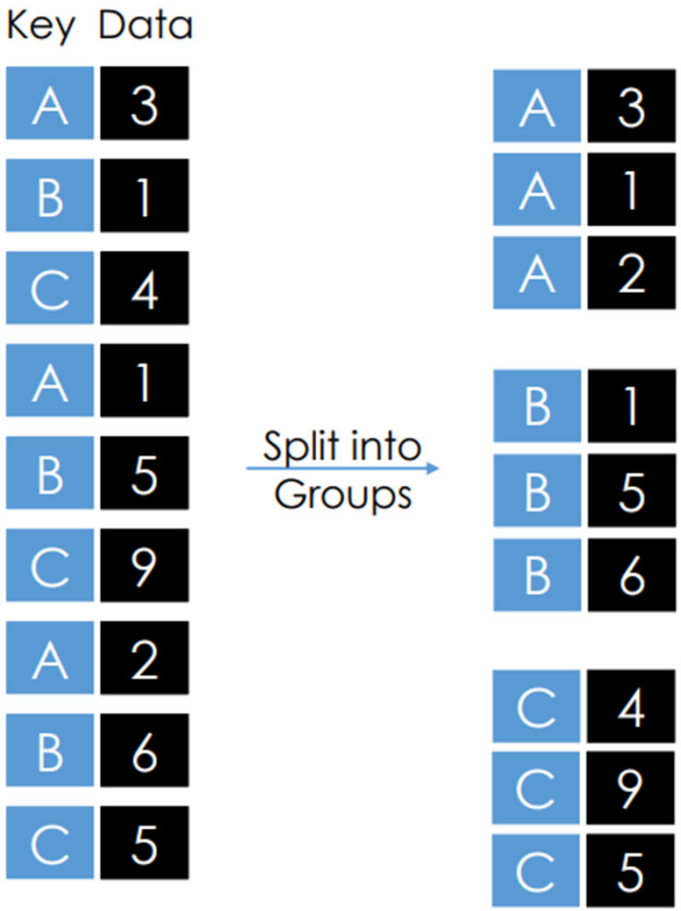
How can we change the granularity of data in a table?

Key	Data
A	3
B	1
C	4
A	1
B	5
C	9
A	2
B	6
C	5

A	3
A	1
A	2

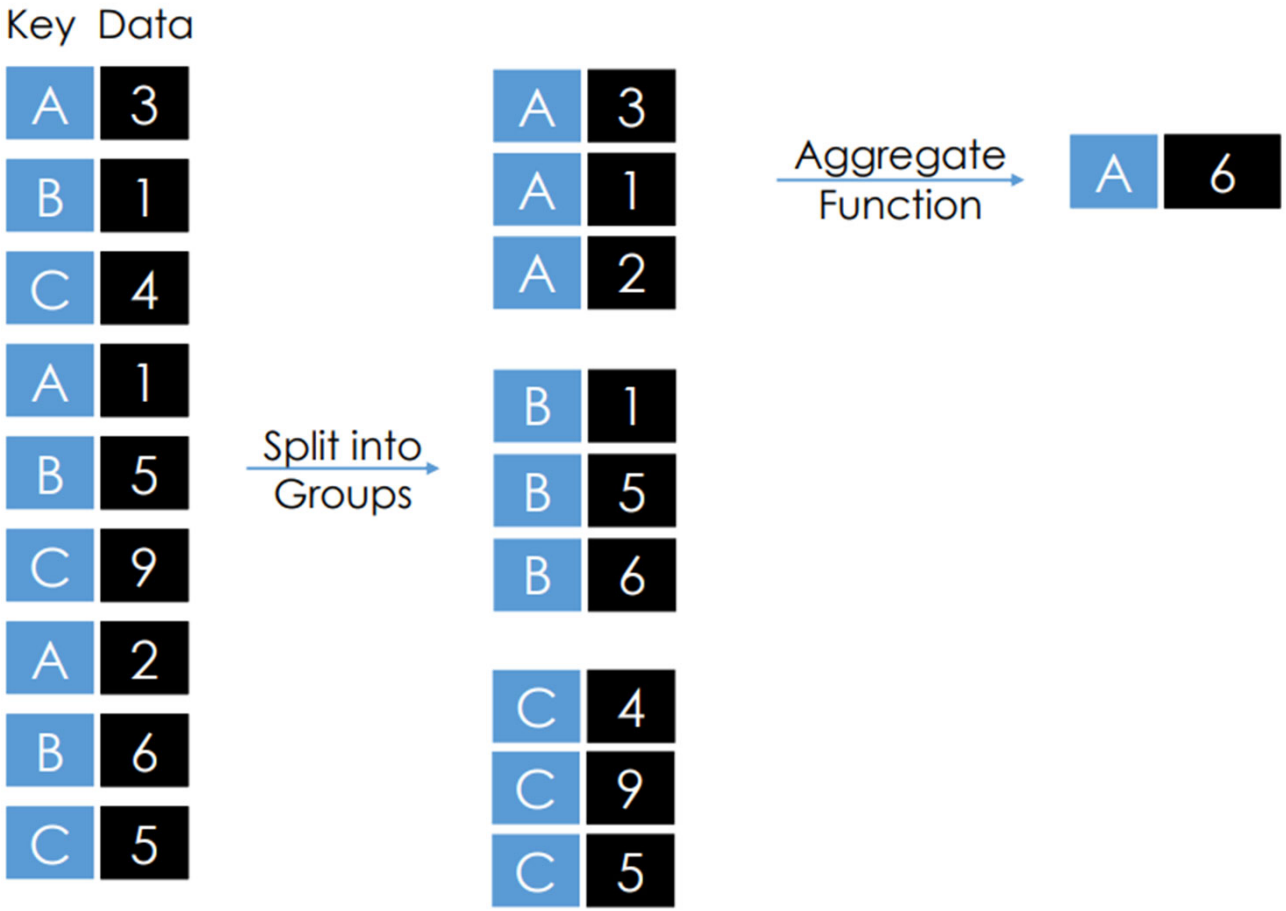
Grouping

How can we change the granularity of data in a table?



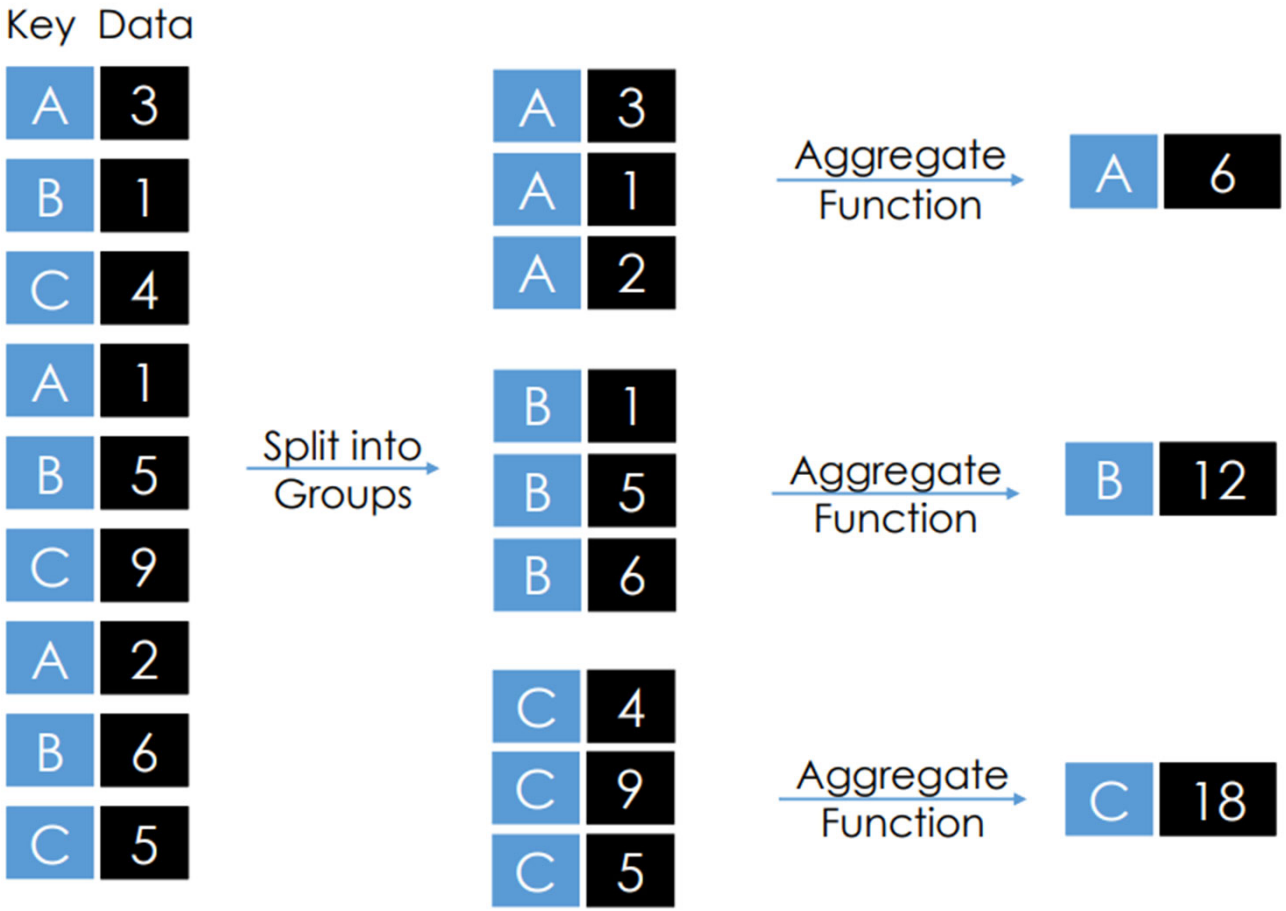
Grouping

How can we change the granularity of data in a table?



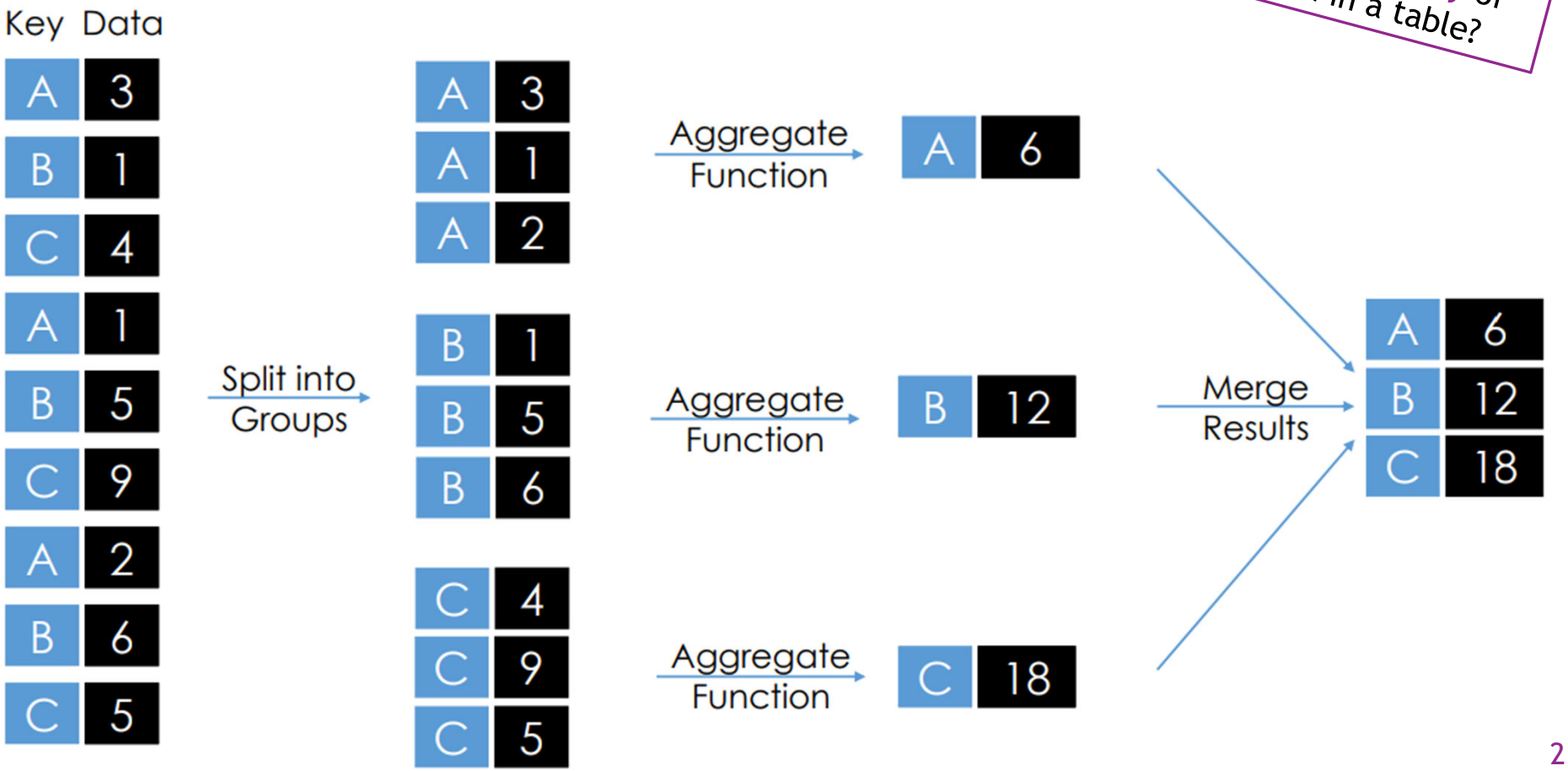
Grouping

How can we change the granularity of data in a table?



Grouping

How can we change the granularity of data in a table?



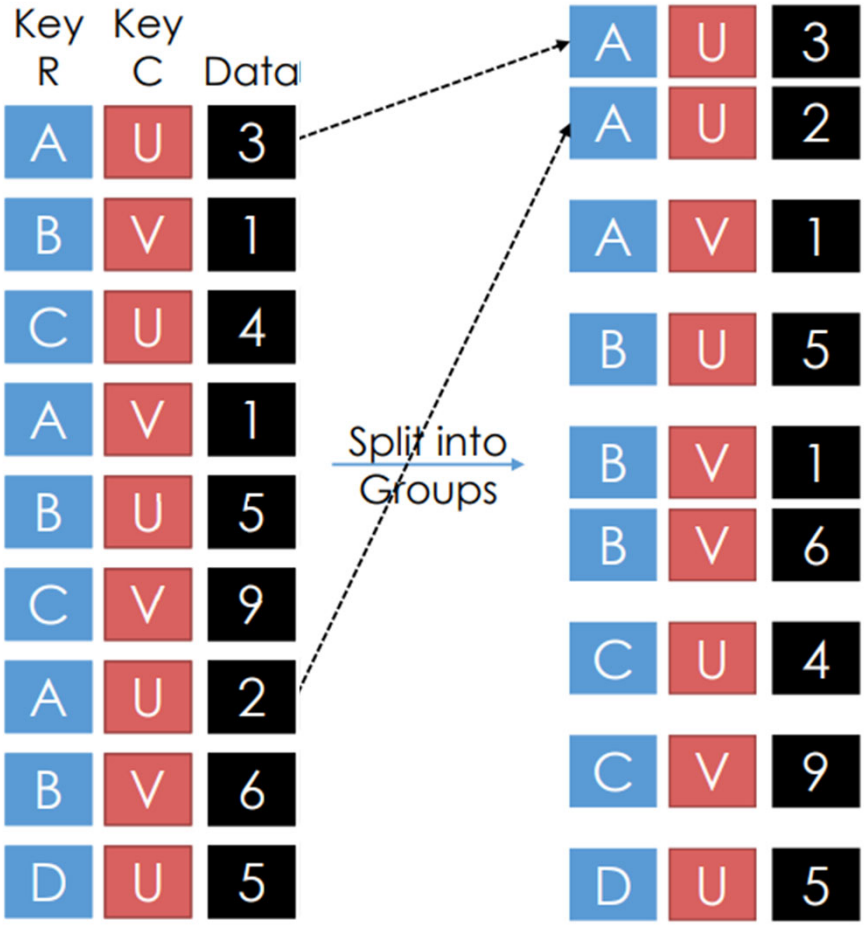
Pivoting

How can we change the granularity of data in a table?

Key R	Key C	Data
A	U	3
B	V	1
C	U	4
A	V	1
B	U	5
C	V	9
A	U	2
B	V	6
D	U	5

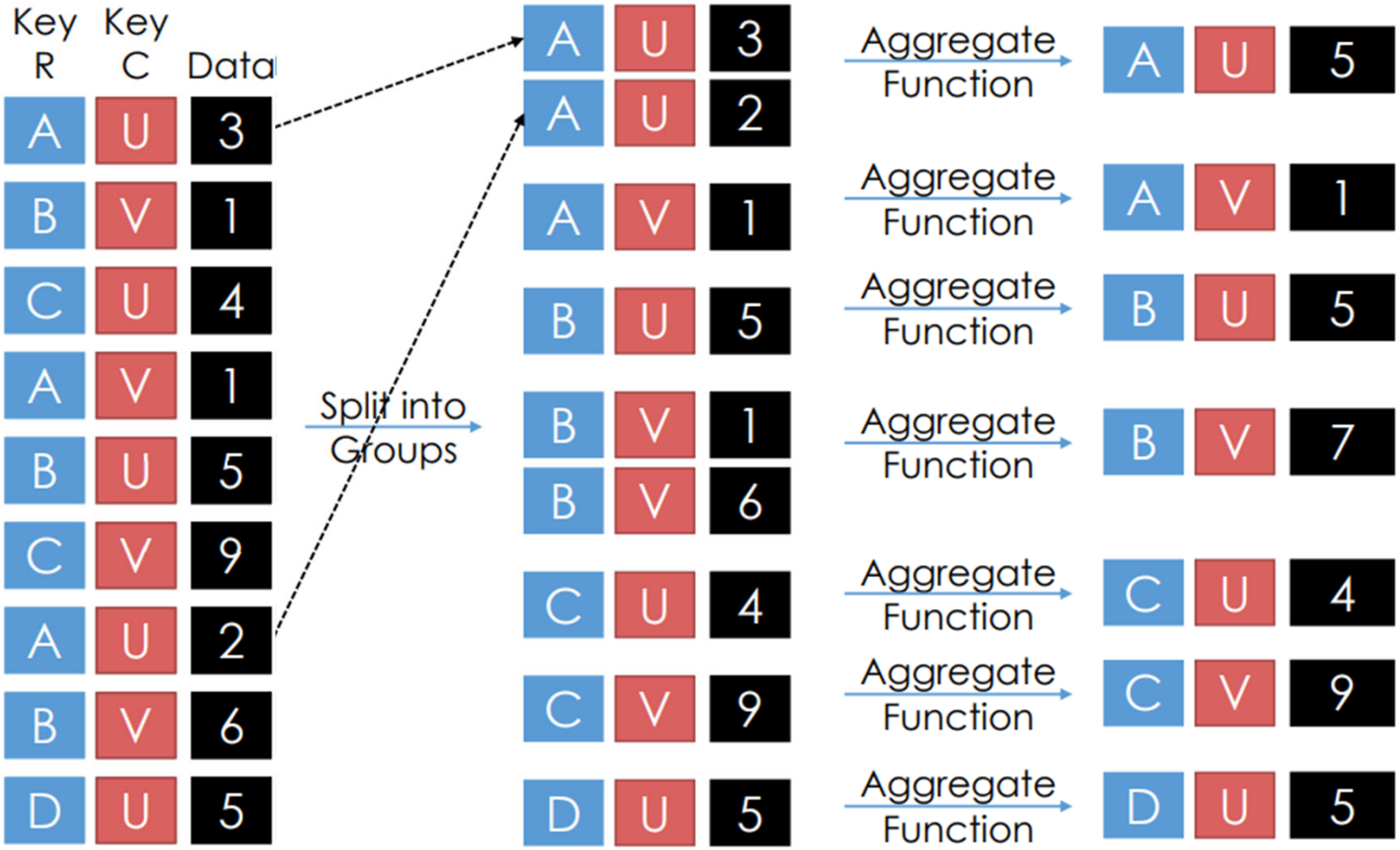
Pivoting

How can we change the granularity of data in a table?

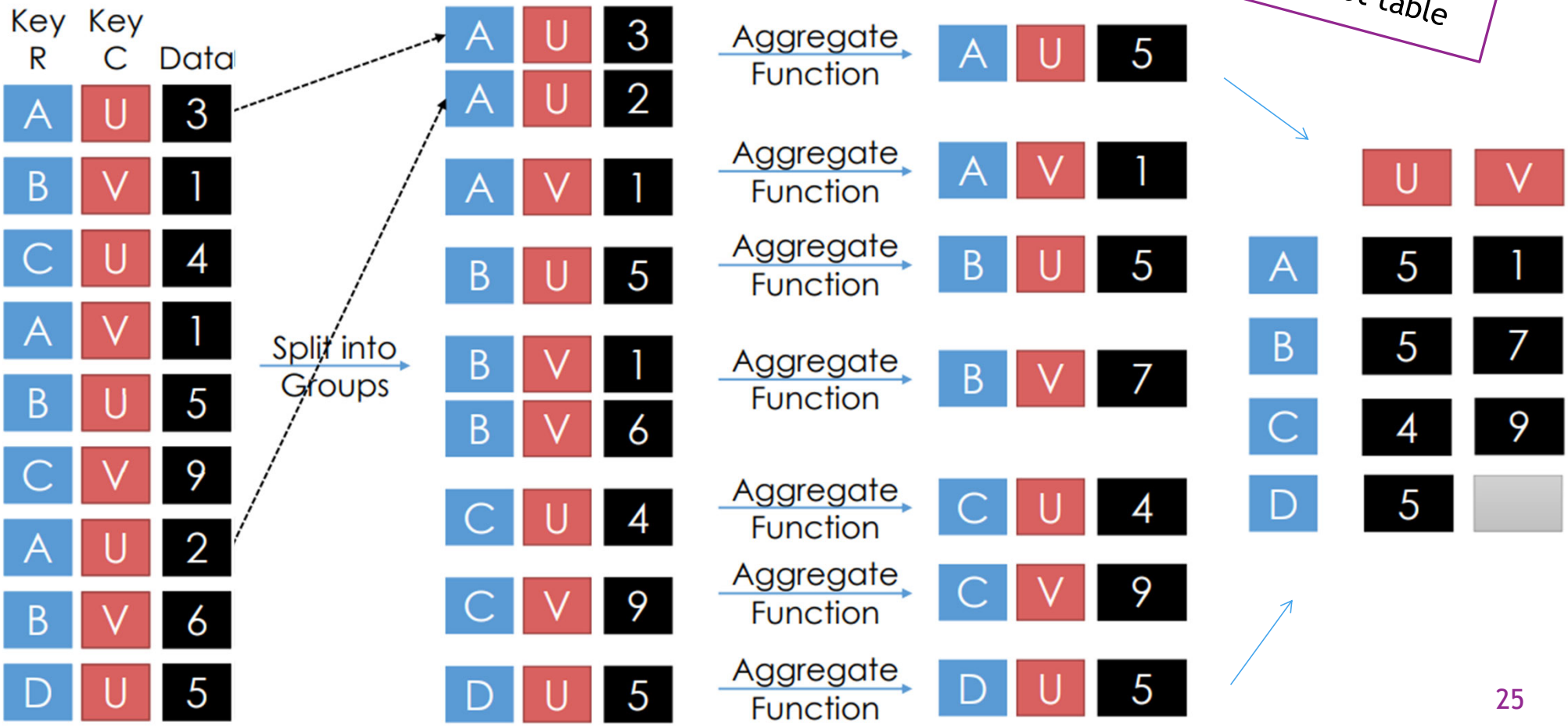


Pivoting

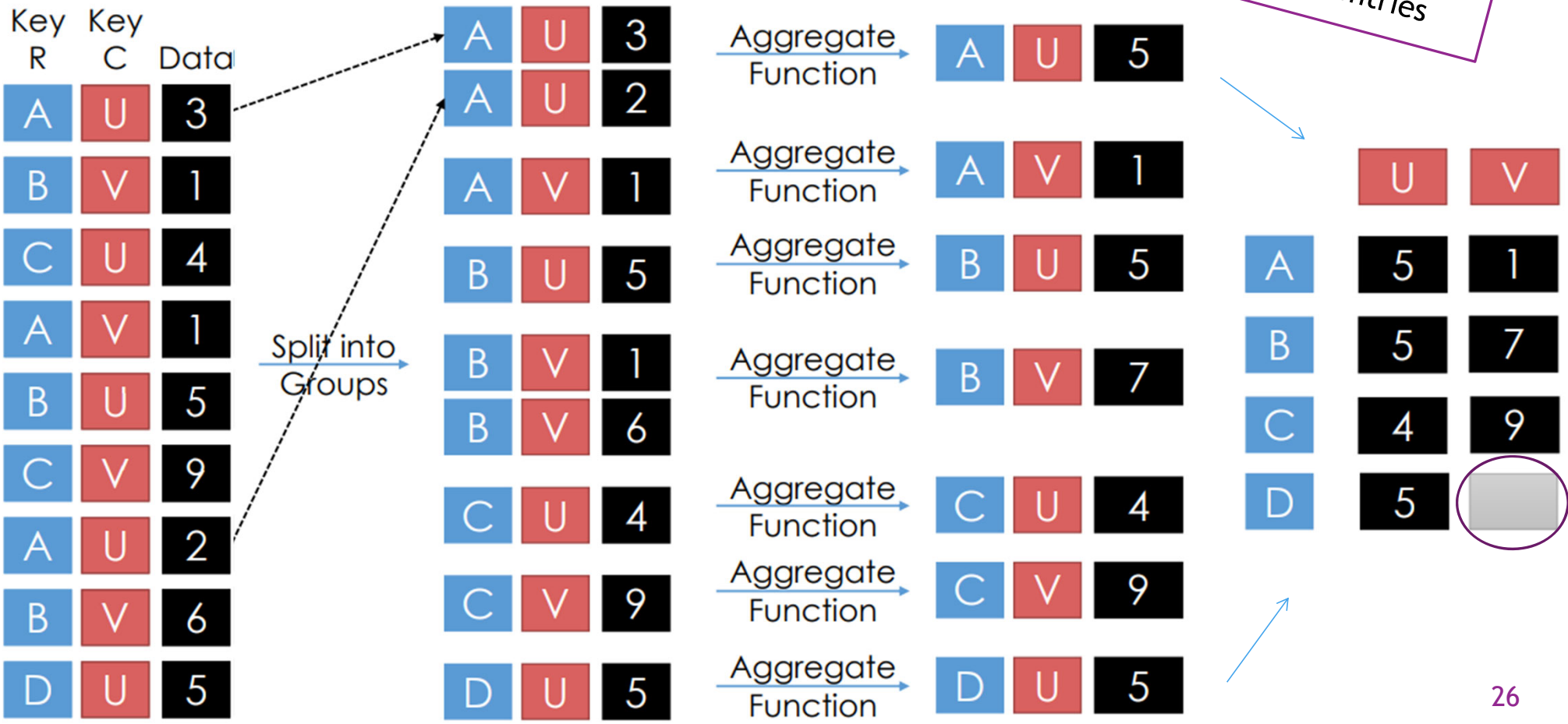
How can we change the granularity of data in a table?



Pivoting


Key R
Key C
Data
A U 3
B V 1
C U 4
A V 1
B U 5
C V 9
A U 2
B V 6
D U 5
Split into Groups
A U 3
A U 2
A V 1
B U 5
B V 1
B V 6
C U 4
C V 9
D U 5
Aggregate Function
A U 5
Aggregate Function
A V 1
Aggregate Function
B U 5
Aggregate Function
B V 7
Aggregate Function
C U 4
Aggregate Function
C V 9
Aggregate Function
D U 5
Unstack the groups into a pivot table
U V
A 5 1
B 5 7
C 4 9
D 5

Pivoting


Key R
Key C
Data
A U 3
B V 1
C U 4
A V 1
B U 5
C V 9
A U 2
B V 6
D U 5
Split into Groups
A U 3
A U 2
A V 1
B U 5
B V 1
B V 6
C U 4
C V 9
D U 5
Aggregate Function
A U 5
Aggregate Function
A V 1
Aggregate Function
B U 5
Aggregate Function
B V 7
Aggregate Function
C U 4
Aggregate Function
C V 9
Aggregate Function
D U 5
Need to handle missing entries
U V
A 5 1
B 5 7
C 4 9
D 5

Summary

- Operations on Tables
  - Inspecting
  - Sorting
  - Summarizing
- Grouping and Pivoting
  - Find the most popular name in New York
  - Find all names that start with E.
  - Sort names by occurrence of dr and ea
  - Find the name whose popularity has changed the most.
  - Count the number of female and male babies born in each year

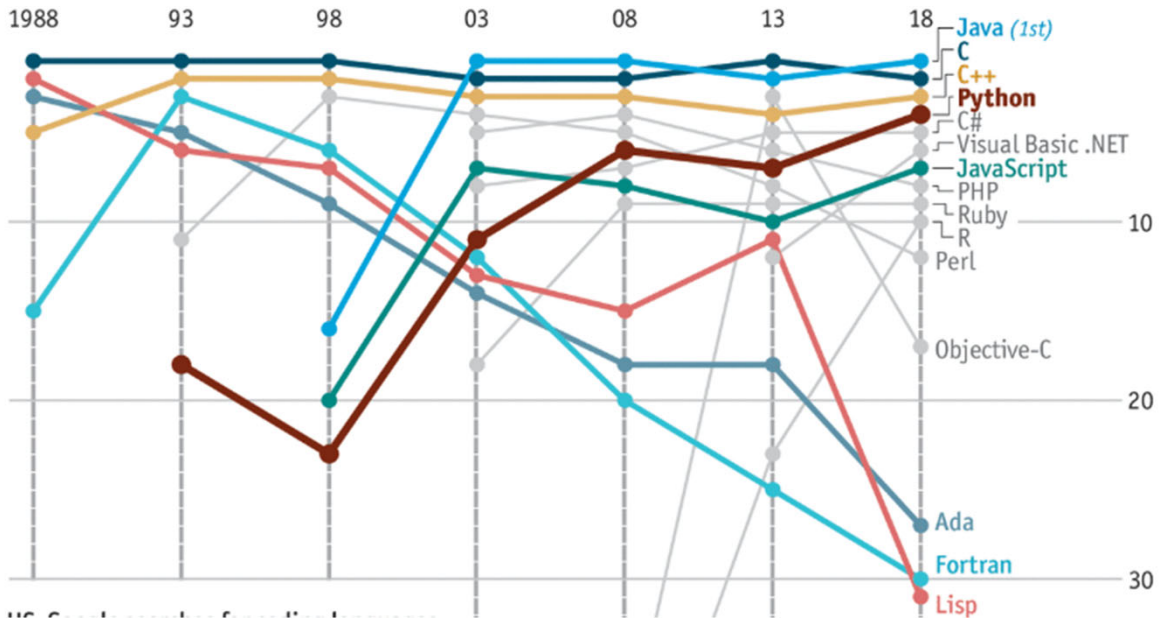
Goals

- Apply
- Group
  - agg
  - size,…
  - filter,
- Pivot

Questions

- Questions on Piazza?
  - Please provide your feedback along with questions
- Question for You!

Why use Python for studying data instead of another programming language?



Questions

- Questions on Piazza?
  - Please provide your feedback along with questions
- Question for You!

Why use Python for studying data instead of another programming language?