NYU Center for Data Science

DS-UA 112
Introduction to Data Science

Lecture 7
Cleaning Data – Split, Apply, Combine

Reminders

- Survey 2

Reminders

- Survey 2
- Homework
  - JupyterHub and Gradescope
  - Grader Contact Information

Reminders

- Survey 2
- Homework
  - JupyterHub and Gradescope
  - Grader Contact Information
- Forum
  - Homework, Lab, Lecture

Reminders

- Survey 2
- Homework
  - JupyterHub and Gradescope
  - Grader Contact Information
- Forum
  - Homework, Lab, Lecture
- Project 1

Agenda

- Review
  - Drop, Fill, Impute

Agenda

- Review
- Lesson
  - JSON, YAML, XML

Agenda

- Review
- Lesson
- Demo
  - Group, Merge

Data Cleaning

Data Cleaning

- File format (e.g. csv vs tsv)
- Encoding and/or formatting of values (e.g., strings vs numbers)
- Missing values (e.g. null, NA, or NaN)

Data Cleaning

- File format (e.g. csv vs tsv)
- Encoding and/or formatting of values (e.g., strings vs numbers)
- Missing values (e.g. null, NA, or NaN)
- Extracting features (e.g., the year from a timestamp)
- Unit conversion for different scales (kilometers vs miles)

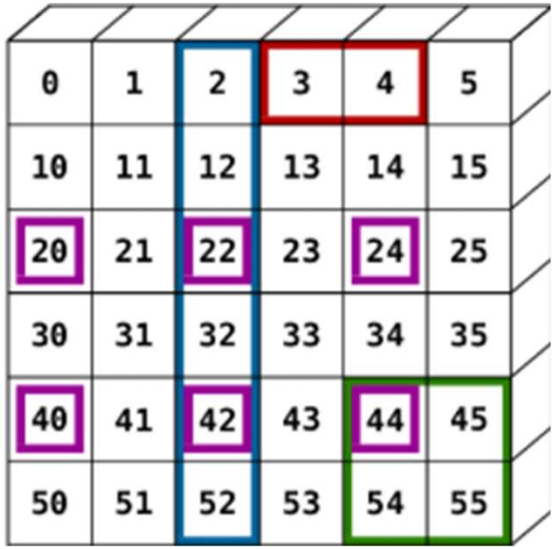
Data Cleaning

Variable	Description	Data Type
bwt	Birth weight in ounces (999 unknown)	Numerical
gestation	Length of pregnancy in days (999 unknown)	Numerical
parity	0= first born, 9=unknown	Nominal
age	mother's age in years	Numerical
height	mother's height in inches (99 unknown)	Numerical
weight	Mother's prepregnancy weight in pounds (999 unknown)	Numerical
smoke	Smoking status of mother (0=not now, 1=yes now, 9=unknown)	Nominal

Arrays

- Fixed
  - Type
  - Size

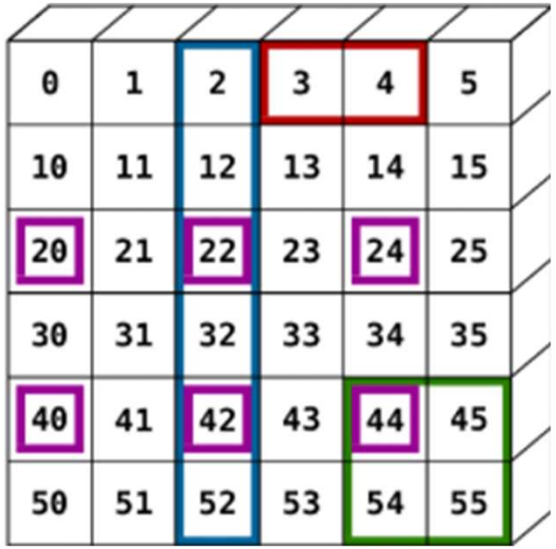
```
>>> a[0, 3:5]
array([3, 4])

>>> a[4:, 4:]
array([[44, 55],
       [54, 55]])

>>> a[:, 2]
a([2, 12, 22, 32, 42, 52])

>>> a[2::2, ::2]
array([[20, 22, 24],
       [40, 42, 44]])
```



Arrays

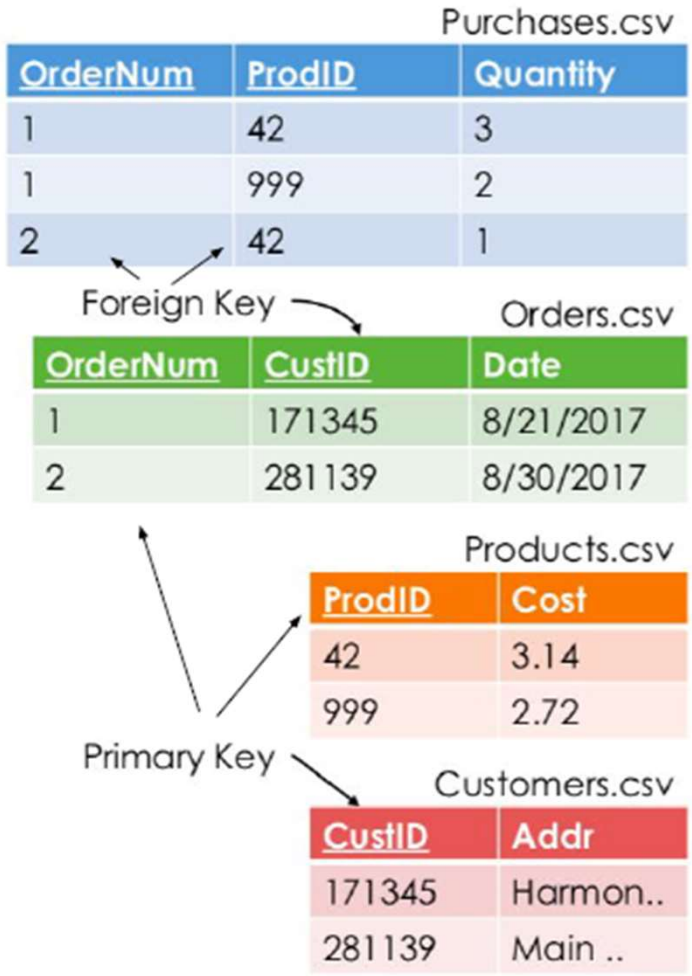
- Fixed
  - Type
  - Size
- Operations
  - Reshaping to match size…called broadcasting
  - Element by element avoiding loops…called vectorization

```
>>> a[0, 3:5]
array([3, 4])

>>> a[4:, 4:]
array([[44, 55],
       [54, 55]])

>>> a[:, 2]
a([2, 12, 22, 32, 42, 52])

>>> a[2::2, ::2]
array([[20, 22, 24],
       [40, 42, 44]])
```



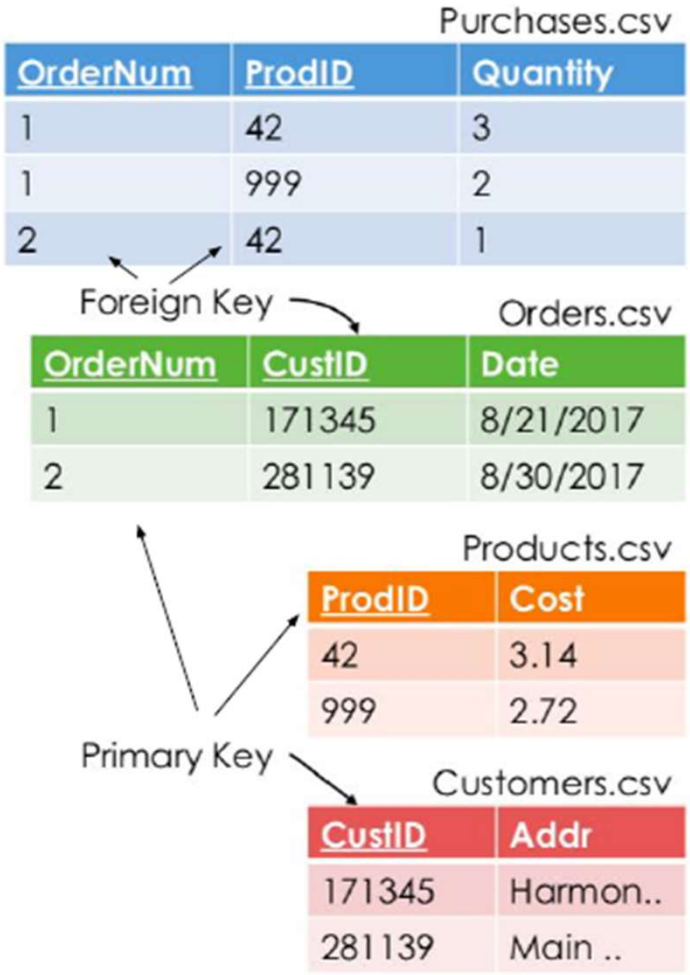
Keys

- Primary Key
  - Set of columns distinguish rows
    - Unique for each row
    - Ensures row can be identified



Keys

- Primary Key
  - Set of columns distinguish rows
    - Unique for each row
    - Ensures row can be identified
- Foreign Key
  - Column with values that are primary keys for other rows.
    - Reference to a row in the same table or a different table.
    - Joining tables expands the reference with values



File Formats

- JSON
- YAML
- XML

File Formats

- JSON
  - Contain data in nested format
  - Resemble Python dictionaries with braces and colons

File Formats

- YAML
  - Stores data in nested format
  - Lacks braces but contains dashes

File Formats

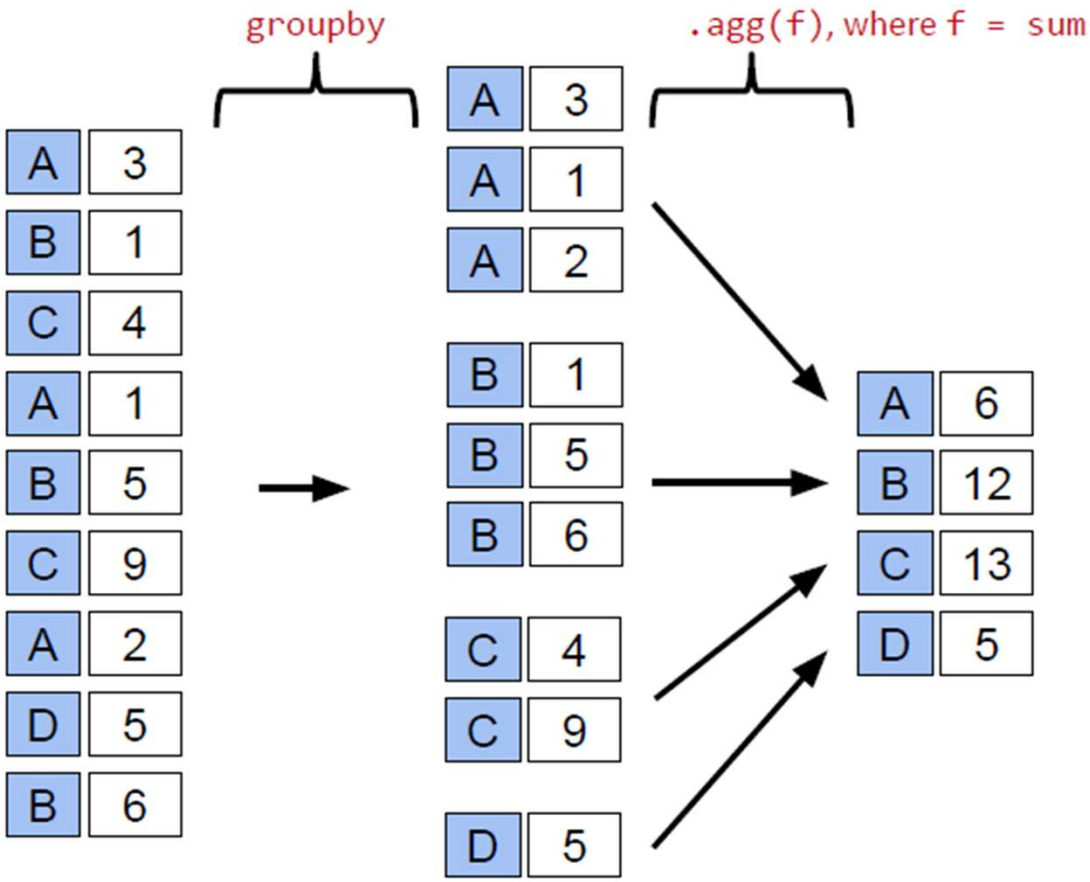
- XML
  - Stores data in tree format
  - Contains Text, Tags and Attributes

File Formats

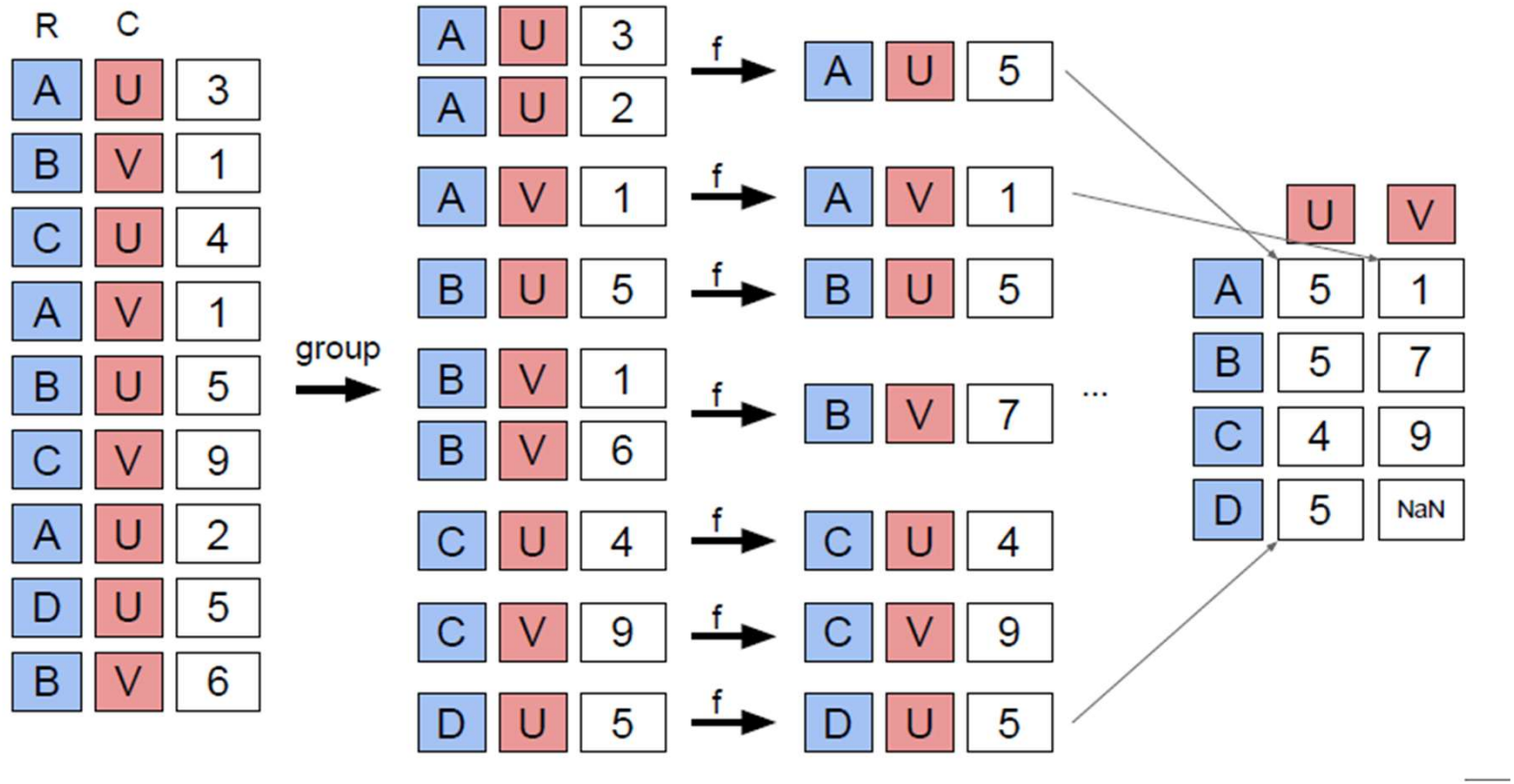
XML	JSON	YAML

<Servers>
<Server>
<name>Server1</name>
<owner>John</owner>
<created>123456</created>
<status>active</status>
</Server>
</Servers>
	{
	Servers: [
	{
	name: Server1,
	owner: John,
	created: 123456,
	status: active
	}
	]
	}
		Servers:
		- name: Server1
		owner: John
		created: 123456
		status: active

Group


groupby
.agg(f), where f = sum
A 3
B 1
C 4
A 1
B 5
C 9
A 2
D 5
B 6
A 3
A 1
A 2
B 1
B 5
B 6
C 4
C 9
D 5
A 6
B 12
C 13
D 5

Pivot


R C
A U 3
B V 1
C U 4
A V 1
B U 5
C V 9
A U 2
D U 5
B V 6
group
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
f
A U 5
A V 1
B U 5
B V 7
C U 4
C V 9
D U 5
...
U V
A 5 1
B 5 7
C 4 9
D 5 NaN

Take-Aways

- What is the format?
  - Tabular data: CSV, TSV
  - Nested data: JSON, YAML, XML

Take-Aways

- What is the format?
  - Tabular data: CSV, TSV
  - Nested data: JSON, YAML, XML
- What are the columns?
  - What is the type of each column?

Take-Aways

- What is the format?
  - Tabular data: CSV, TSV
  - Nested data: JSON, YAML, XML
- What are the columns?
  - What is the type of each column?
- Does the data reference other data?
  - Can we join the data?