Data Cleaning (Reading: 5.1-5.3)

Learning goals:

- Understand common cleaning requirements.
- Learn some challenges that missing values present.
- Learn how to join tables in pandas.

UC Berkeley Data 100 Summer 2019
Sam Lau

(Slides adapted from John DeNero, Josh Hug, Joey Gonzalez, Deb Nolan, Fernando Perez, & Joe Hellerstein)

Announcements

- There is a live lecture Piazza thread: Raguvir will post.
- Exam Conflict form link changed: http://bit.ly/su19-alt-final
    - Due Friday at 11:59pm
    - DSP exams will have a separate form, will send later
- Small group tutoring is starting next week; more info soon
- Project 1 due next Tues.
- DSP students: I don’t have access to your letters. Please email them to me.

Data Cleaning

Big Data Borat
@BigDataBorat

Following

In Data Science, 80% of time spent prepare data,

What is Data Cleaning?

The process of transforming raw data to facilitate subsequent analysis.

Data cleaning often addresses:

- Structure of the data
- Encoding and/or formatting of values (e.g., strings vs numbers)
- Missing and/or corrupted values
- Unit conversion and interpretation of magnitudes
- Extracting features of complex values (e.g., the year from a timestamp)

What is Data Cleaning?

Data cleaning is a time-consuming but critical part of DS

Data cleaning is often a reaction to other stages of the data science lifecycle:

- Exploratory data analysis identifies issues with the raw data.
- Surprising inferential results might be due to misinterpreted or bad data.
- New hypotheses or questions prompt an investigation some aspect of data.

Bad Data

All of these are commonly seen in the real-world:

- Zeros replace missing values
- Spelling inconsistent (esp with human-entered data)
- Rows are duplicated
- Inconsistent date formats (e.g. 10/9/15 vs. 9/10/15)
- Units not specified

https://github.com/Quartz/bad-data-guide

From the Quartz Bad Data Guide:

“There is no worse way to screw up data than to let a single human type it in, without validation.”

“For example, I once acquired the complete dog licensing database for Cook County, Illinois.”

“...this database contained at least 250 spellings of Chihuahua. Even with the best tools available, data this messy can't be saved. They are effectively meaningless.“

Rectangular Data

Rectangular data sets are easy to manipulate, visualize, and combine.

Tables (DataFrames):

- Each labeled column has values of the same type.
- Manipulated using group, sort, join, etc.
- Formal description of data transformations is called *relational algebra*.

Rectangular Data

Matrices:

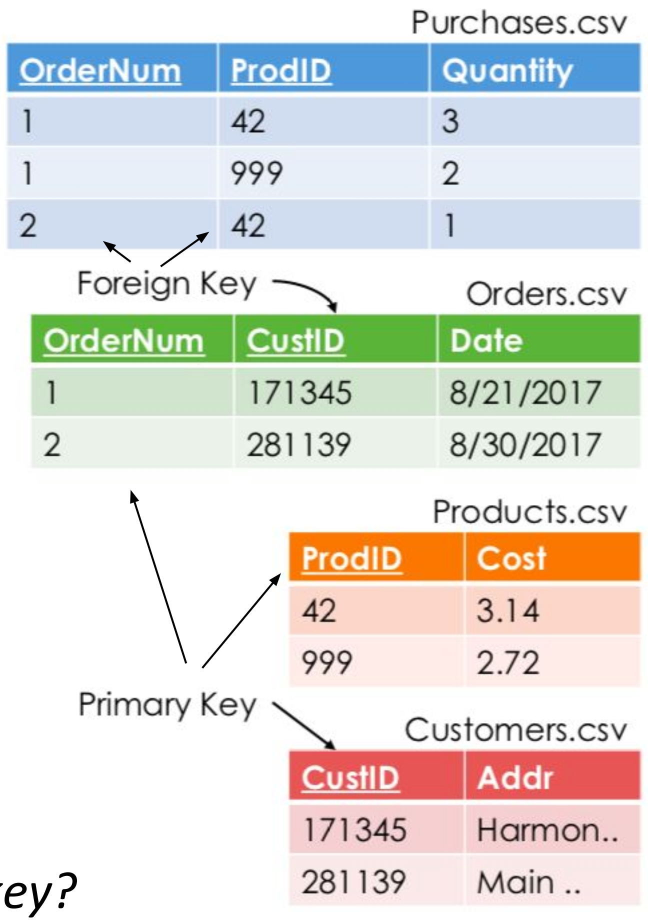
- **All values have the same type.**
- Manipulated using multiplication, addition, and element-wise operations.
- Most useful manipulation is linear, described by *linear algebra*.

A lot of data cleaning effort is spent making data more rectangular.

Keys

Primary key: the column (or set of columns) that determines the values in the remaining columns.

- Unique for each row & 1-to-1 with entities.
- Ensures that the row can be identified, even after appending more data. E.g., SSN



Purchases.csv

OrderNum	ProdID	Quantity
1	42	3
1	999	2
2	42	1

Orders.csv

OrderNum	CustID	Date
1	171345	8/21/2017
2	281139	8/30/2017

Products.csv

ProdID	Cost
42	3.14
999	2.72

Customers.csv

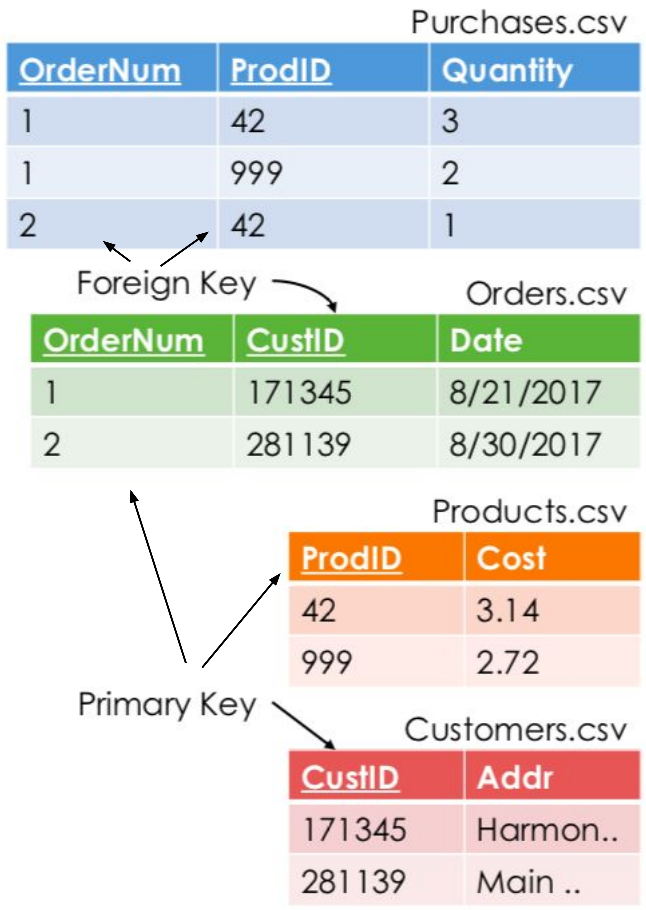
CustID	Addr
171345	Harmon..
281139	Main ..

Is an email address a good primary key?

Keys

Foreign key: a column containing values that are primary keys for other rows.

- A foreign key serves as a reference to a row.
- Joining tables expands the reference with values from the referenced row.
- The referenced row can be in the same table or a different table.



Purchases.csv

OrderNum	ProdID	Quantity
1	42	3
1	999	2
2	42	1

Orders.csv

OrderNum	CustID	Date
1	171345	8/21/2017
2	281139	8/30/2017

Products.csv

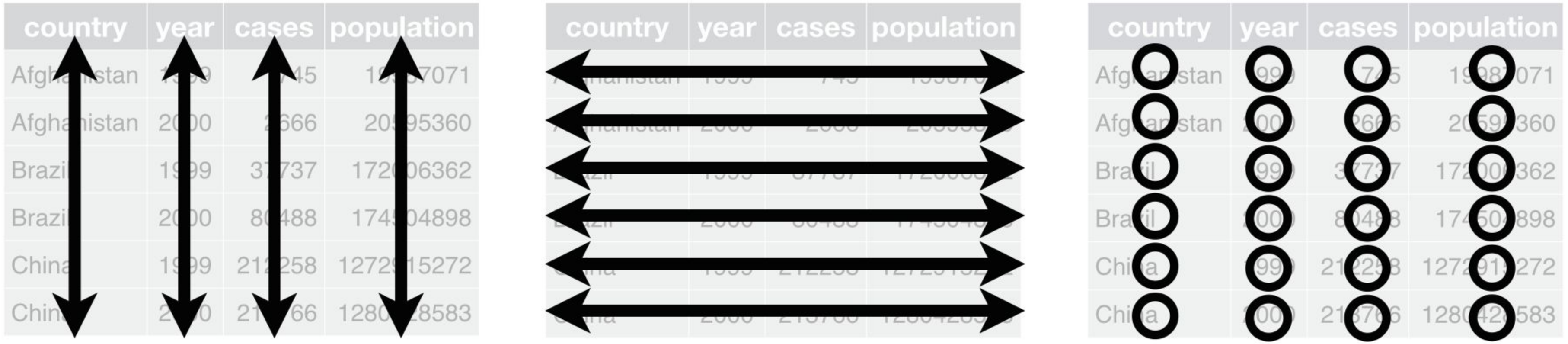
ProdID	Cost
42	3.14
999	2.72

Customers.csv

CustID	Addr
171345	Harmon..
281139	Main ..

Tidy Data

- Rectangular dataset with a particularly useful format:
  - Every variable has its own column
  - Every observation has its own row
  - Every value has its own cell



variables observations values

Tidy Data

- “Tidy datasets are all alike, but every messy dataset is messy in its own way.” –– Hadley Wickham

Tidy

	country	year	cases	population
0	Afghanistan	1999	745	19987071
1	Afghanistan	2000	2666	20595360
2	Brazil	1999	37737	172006362
3	Brazil	2000	80488	174504898
4	China	1999	212258	1272915272
5	China	2000	213766	1280428583

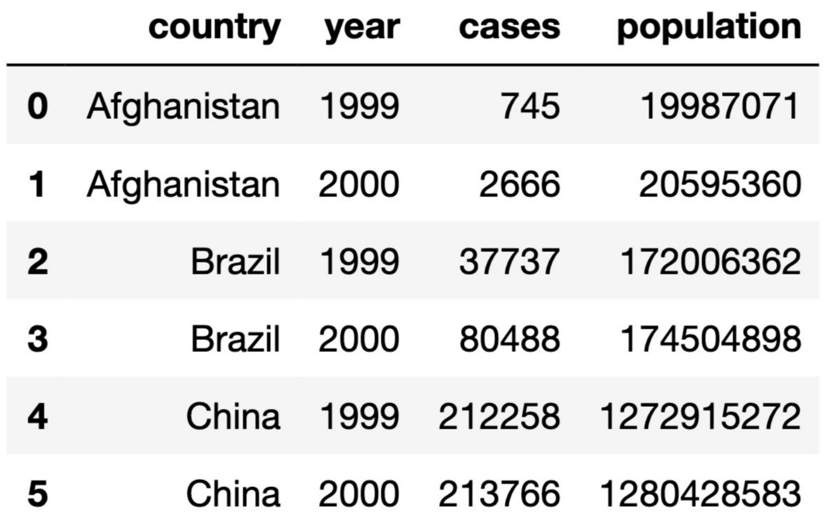
Not tidy!

	country	year	type	count
0	Afghanistan	1999	cases	745
1	Afghanistan	1999	population	19987071
2	Afghanistan	2000	cases	2666
3	Afghanistan	2000	population	20595360
4	Brazil	1999	cases	37737
5	Brazil	1999	population	172006362

Tidy Data

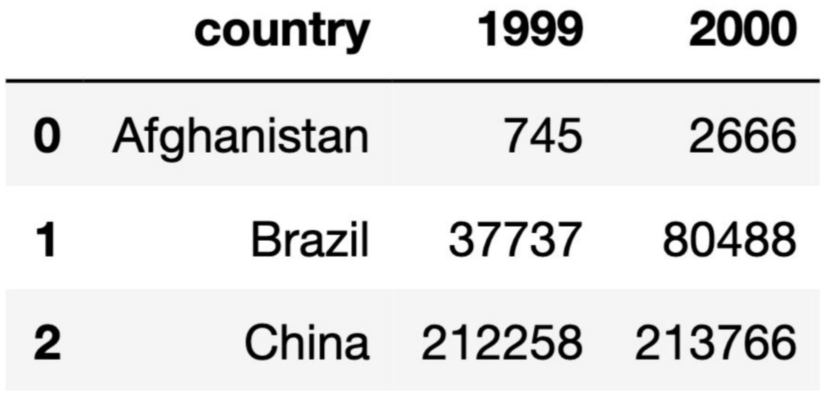
- “Tidy datasets are all alike, but every messy dataset is messy in its own way.” –– Hadley Wickham

Tidy



	country	year	cases	population
0	Afghanistan	1999	745	19987071
1	Afghanistan	2000	2666	20595360
2	Brazil	1999	37737	172006362
3	Brazil	2000	80488	174504898
4	China	1999	212258	1272915272
5	China	2000	213766	1280428583

Not tidy!



	country	1999	2000
0	Afghanistan	745	2666
1	Brazil	37737	80488
2	China	212258	213766

(Tidy data demo)

Data File Formats

Comma-Separated Values

pandas.pydata.org/pandas-docs/stable/reference/api/pandas.read_csv.html

```
pandas.read_csv(filepath_or_buffer, sep=', ', delimiter=None, header='infer',
names=None, index_col=None, usecols=None, squeeze=False, prefix=None,
mangle_dupe_cols=True, dtype=None, engine=None, converters=None, true_values=None,
false_values=None, skipinitialspace=False, skiprows=None, skipfooter=0, nrows=None,
na_values=None, keep_default_na=True, na_filter=True, verbose=False,
skip_blank_lines=True, parse_dates=False, infer_datetime_format=False,
keep_date_col=False, date_parser=None, dayfirst=False, iterator=False,
chunksize=None, compression='infer', thousands=None, decimal=b'.',
lineterminator=None, quotechar='"', quoting=0, doublequote=True, escapechar=None,
comment=None, encoding=None, dialect=None, tupleize_cols=None,
error_bad_lines=True, warn_bad_lines=True, delim_whitespace=False, low_memory=True,
memory_map=False, float_precision=None)
```

Reading CSV Files

Sometimes the first line contains column labels:

`header='infer'` (default)

Other times you need to provide your own labels:

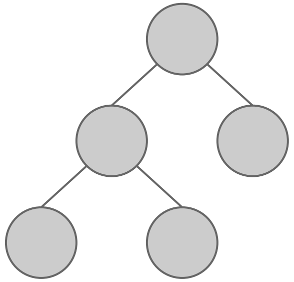
`header=None`

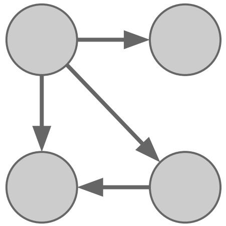
- Reasonable labels come from inspecting the data or reading the docs.
- Don't accidentally skip the first row of data by interpreting it as a header.

JSON, XML, HTML, YAML, Protocol Buffers, etc.

There are many formats to represent structured, hierarchical data.

- Most formats consist of values, lists, and dictionaries.
- Dictionaries are typically keyed by strings, but sometimes by integers.
- Structures vary: a list of records, a list of columns, a tree of documents, etc.





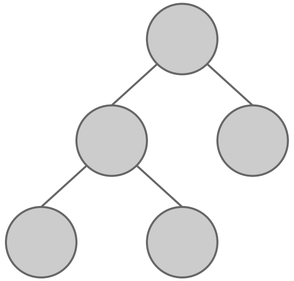
JSON, XML, HTML, YAML, Protocol Buffers, etc.

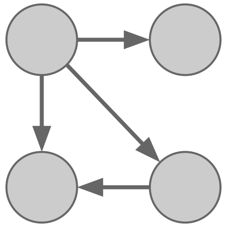
Converting hierarchical data to rectangular data often involves keys. To represent…

- A tree: one column for the node ID, one for the parent ID
- A graph: one table for nodes; one table for edges

In general, traverse the data and yield lists of values (rows).

(Demo)





Log Files

```
169.237.46.168 - - [26/Jan/2014:10:47:58 -0800] "GET
/stat141/Winter04 HTTP/1.1" 301 328
"http://anson.ucdavis.edu/courses/"  "Mozilla/4.0 (compatible; MSIE
6.0; Windows NT 5.0; .NET CLR 1.1.4322)”
```

```
169.237.6.168 - - [8/Jan/2014:10:47:58 -0800] "GET
/stat141/Winter04/ HTTP/1.1" 200 2585
"http://anson.ucdavis.edu/courses/" "Mozilla/4.0 (compatible; MSIE
6.0; Windows NT 5.0; .NET CLR 1.1.4322)"
```

Break!
Fill out Attendance:
http://bit.ly/at-d100

Missing Values

What to do with Missing Values

Often, rectangular data sets have missing values:

- Field lost, hidden, removed, replaced, or never entered.
- Or, perhaps the entity described by a record does not have a particular attribute.
  E.g., some people don’t have a permanent address.

What to do with Missing Values

How to treat missing values depends on goals and context.

Discarding whole records (because of missing values) results in a sample.

- That sample isn't a random sample, so statistical inference is inappropriate.
- The sample will often be biased — not representative of the population.

What to do with Missing Values

Replacing missing values within records has consequences for analysis.

- If possible, replace missing values with the true value that was removed.
- When imputing values, document the change and consider the effect.
  E.g., replacing missing values with the mean will decrease the variance.

(Demo)

Joins

(Demo)

Joins

Will usually want to join a foreign key of one table with the primary key of another.

Use `pd.merge(...)` .

- `pd.join(...)` is a shorthand for `pd.merge(...)` but only works when joining tables using their indices.

Summary

- Reality is cruel — many different formats of data, quality ranges widely.
- Data cleaning tackles some problems that arise from data collection.
    - Not all problems can be addressed via cleaning! You can't "clean" a non-random sample into a random one.
- Goal in this class: Process data so that further analysis only needs to work with with one (or a few) tidy DataFrame(s).