LECTURE 4

Data Wrangling and EDA, Part I

Exploratory Data Analysis and its role in the data science lifecycle

Interlude

Real World Data Scientist:
Timnit Gebru

- Gender Shades
  (with Joy Buolamwini, 2300+ citations)
- **Datasheets for Datasets (580+ citations)**

> The **characteristics of these datasets fundamentally influence a model's behavior**: a model is unlikely to perform well in the wild if its deployment context does not match its training or evaluation datasets, or if these **datasets reflect unwanted societal biases**.

We need EDA because many datasets don't come with comprehensive datasheets/codebooks!

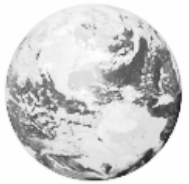
Just finished…

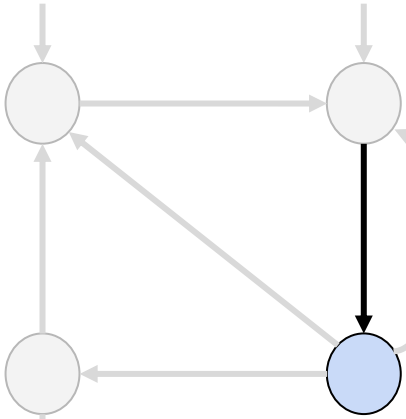
http://abcnews.go.com/Lifestyle/silly-baby-panda-falls-flat-face-public-debut/story?id=42481478

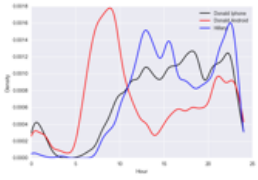
The Next Step

EDA Guiding Principles

Plan for first few weeks







(Weeks 1 and 2)

Exploring and Cleaning Tabular Data
From `datascience` to `pandas`

(Weeks 2 and 3)

Data Science in Practice
EDA, Data Cleaning, Text processing (regular expressions)

Exploring Tabular Data

Exploring Tabular Data

Record Granularity

Variable Types

Multiple Files

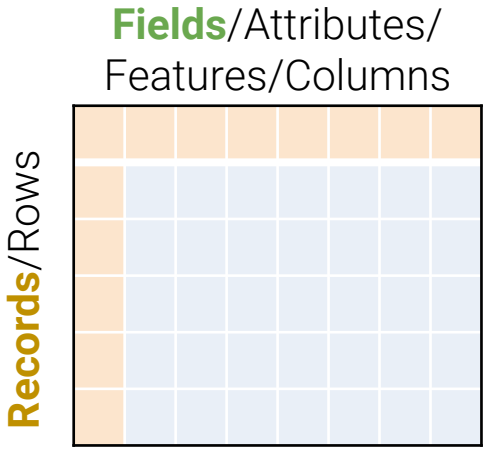
More EDA/Wrangling

Rectangular Data

We prefer rectangular data for data analysis (why?)

- Regular structures are easy manipulate and analyze
- A big part of data cleaning is about transforming data to be more rectangular

Two kinds of rectangular data: **Tables** and **Matrices**.



Tables (a.k.a. dataframes in R/Python and relations in SQL)

- Named columns with different types
- Manipulated using data transformation languages (map, filter, group by, join, …)

Matrices

- Numeric data of the same type (float, int, etc.)
- Manipulated using linear algebra

What are the differences?
Why would you use one over the other?

Tuberculosis – United States, 2021

Summary

What is already known about this topic?

The number of reported U.S. tuberculosis (TB) cases decreased sharply in 2020, possibly related to multiple factors associated with the COVID-19 pandemic.

What is added by this report?

Reported TB incidence (cases per 100,000 persons) increased 9.4%, from 2.2 during 2020 to 2.4 during 2021 but was lower than incidence during 2019 (2.7). Increases occurred among both U.S.-born and non–U.S.-born persons.

What are the implications for public health practice?

Factors contributing to changes in reported TB during 2020–2021 likely include an actual reduction in TB incidence as well as delayed or missed TB diagnoses. Timely evaluation and treatment of TB and latent tuberculosis infection remain critical to achieving U.S. TB elimination.

CDC Morbidity and Mortality Weekly Report (MMWR) 03/25/2022.

What is **incidence**? Why use it here?

How was "9.4% increase" computed?

Question: Can we **reproduce** these rates using government data?

CDC: **source**, 2021

Demo Slides

CSV: Comma-Separated Values

Tuberculosis in the US [CDC source].

CSV is a very common **tabular file format**.

- **Records** (rows) are delimited by a newline: `'\n'`, `"\r\n"`
- **Fields** (columns) are delimited by commas: `','`

Pandas: `pd.read_csv(header=...)`

Fields/Attributes/Features/Columns

Records/Rows

	U.S. jurisdiction	TB cases 2019	...
0	Total	8,900	...
1	Alabama	87	...

Record Granularity

Exploring Tabular Data

Record Granularity

Variable Types

Multiple Files

More EDA/Wrangling

Key Data Properties to Consider in EDA

(we'll come back to this)

Structure -- the "shape" of a data file

Granularity -- how fine/coarse is each datum

Scope -- how (in)complete is the data

Temporality -- how is the data situated in time

Faithfulness -- how well does the data capture "reality"

Granularity: How fine/coarse is each datum?

What does each **record** represent?

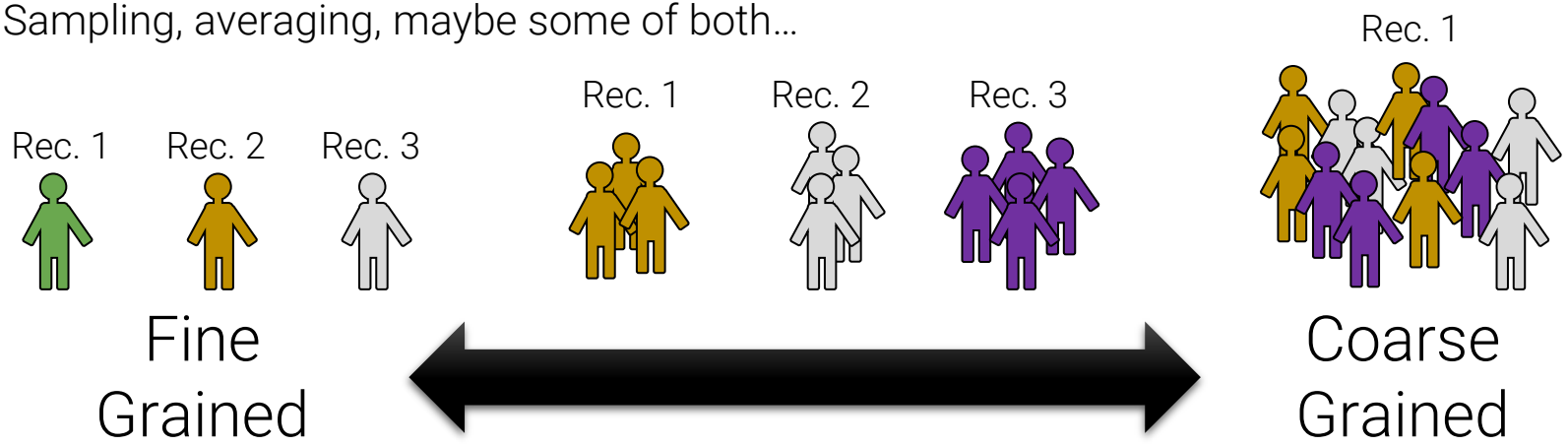
- Examples: a purchase, a person, a group of users

Do all records capture granularity at the same level?

- Some data will include summaries (aka **rollups**) as records.

If the data are **coarse**, how were the records aggregated?

- Sampling, averaging, maybe some of both…



To the demo!!

Variable Types

Exploring Tabular Data

Record Granularity

Variable Types

Multiple Files

More EDA/Wrangling

(we're back to this)

Variable Type

Structure -- the "shape" of a data file

Granularity -- how fine/coarse is each datum

Scope -- how (in)complete is the data

Temporality -- how is the data situated in time

Faithfulness -- how well does the data capture "reality"

Variables are Columns

Let's look at records with the same granularity.

What does each **column** represent?

A **variable** is a **measurement** of a particular concept.

	U.S. jurisdiction	TB cases 2019	…
1	Alabama	87	…
2	Alaska	58	…
…	…	…	…

The U.S. Jurisdiction **variable**

It has two common properties:

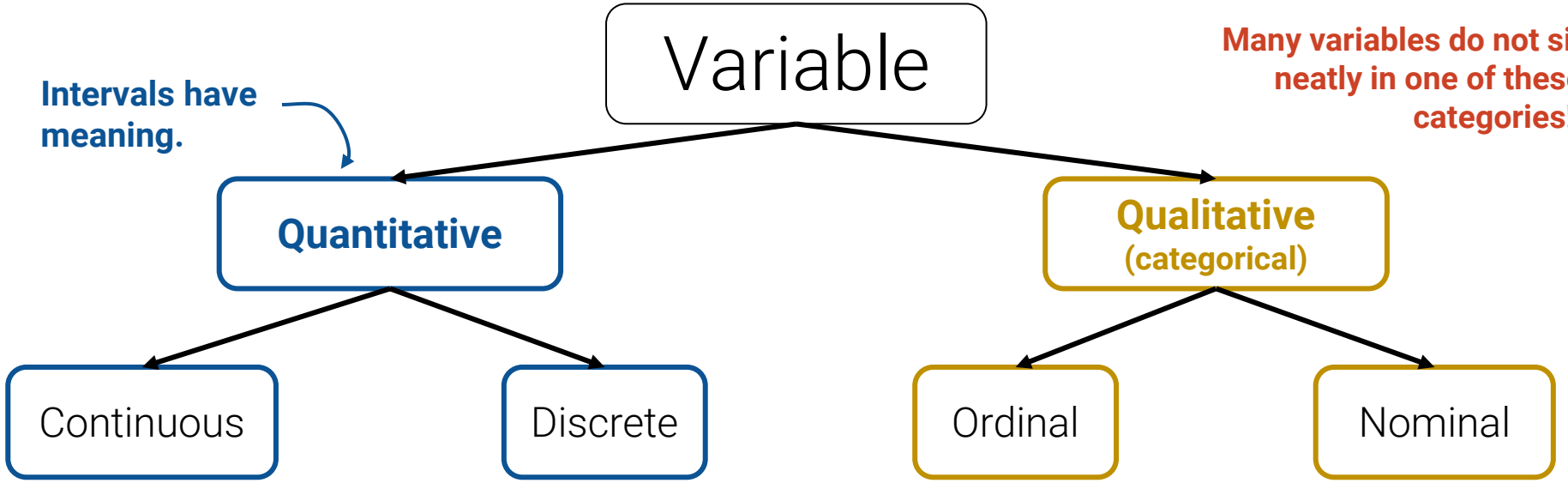
- **Datatype/Storage type**:
  How each variable value is stored in memory. `df[colname].dtype`
  - integer, floating point, boolean, object (string-like), etc.

  Affects which pandas functions you use.
- **Variable type/Feature type**:
  Conceptualized measurement of information (and therefore what values it can take on).
  - Use expert knowledge
  - Explore data itself
  - Consult data codebook (if it exists).

  Affects how you visualize and interpret the data.

⚠ In this class, "variable types" are conceptual!!

Variable Feature Types



Could be measured to arbitrary precision.

Examples:
- Price
- Temperature

Finite possible values

Examples:
- Number of siblings
- Yrs of education

Categories w/ordered levels; no consistent meaning to difference

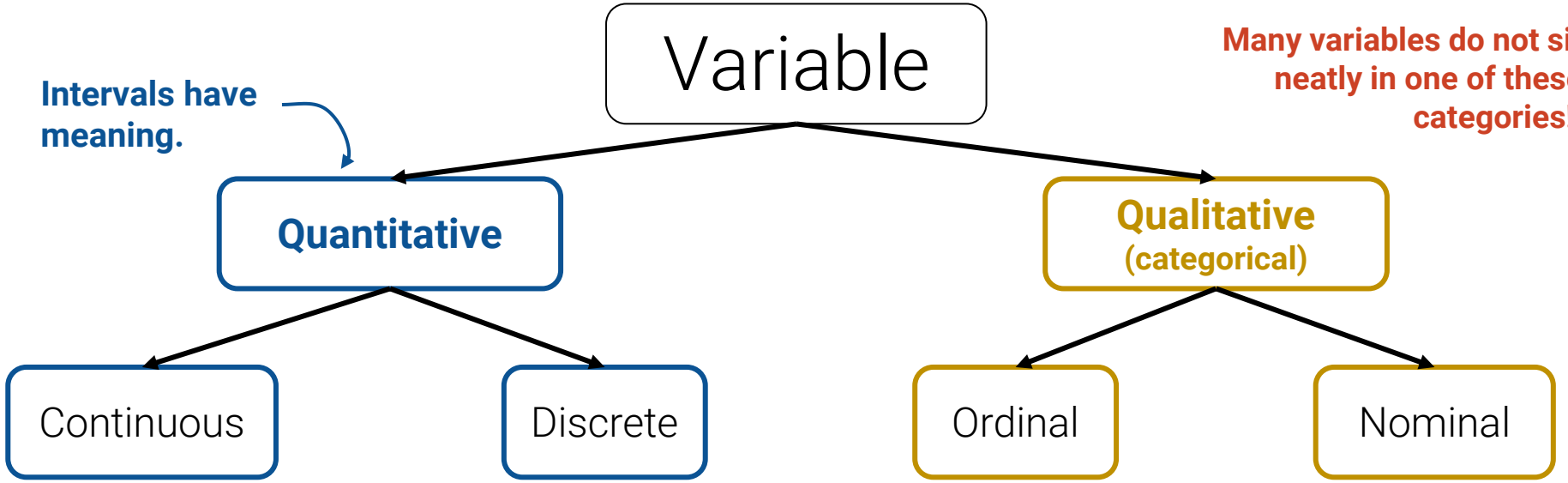
Examples:
- Preferences
- Level of education

Categories w/ no specific ordering.

Examples:
- Political Affiliation
- Cal ID number

Note that **qualitative variables** could have numeric levels; conversely, **quantitative variables** could be stored as strings!

Variable Feature Types



Could be measured to arbitrary precision.

Examples:
- Price
- Temperature

Finite possible values

Examples:
- Number of siblings
- Yrs of education

Categories w/ordered levels; no consistent meaning to difference

Examples:
- Preferences
- Level of education

Categories w/ no specific ordering.

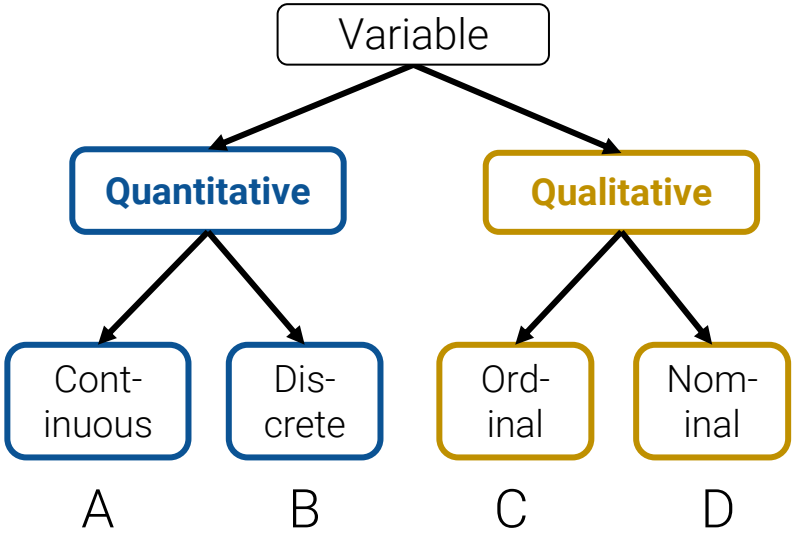
Examples:
- Political Affiliation
- Cal ID number

Note that **qualitative variables** could have numeric levels; conversely, **quantitative variables** could be stored as strings!

Variable Types

What is the feature type (i.e., variable type) of each variable?

Q	Variable	Feature Type
1	CO_2 level (ppm)	
2	Number of siblings	
3	GPA	
4	Income bracket (low, med, high)	
5	Race/Ethnicity	
6	Number of years of education	
7	Yelp Rating	



slido

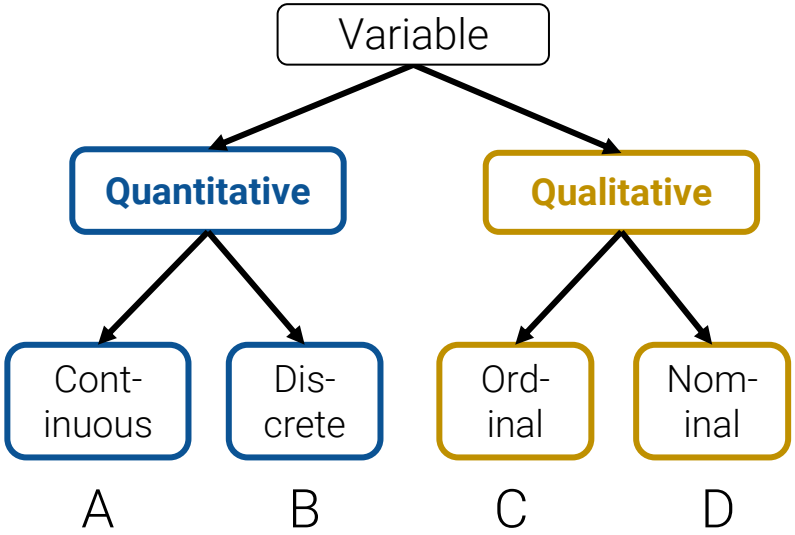
Variable Types

ⓘ Start presenting to display the poll results on this slide.

Variable Types

What is the feature type of each variable?

Q	Variable	Feature Type
1	CO_2 level (ppm)	**A. Quantitative Cont.**
2	Number of siblings	**B. Quantitative Discrete**
3	GPA	**A. Quantitative Cont.**
4	Income bracket (low, med, high)	**C. Qualitative Ordinal**
5	Race/Ethnicity	**D. Qualitative Nominal**
6	Number of years of education	**B. Quantitative Discrete**
7	Yelp Rating	**C. Qualitative Ordinal**



Many of these examples show how "shaggy" these categories are!!
We will revisit variable types when we learn how to visualize variables.

Multiple Files

Exploring Tabular Data

Record Granularity

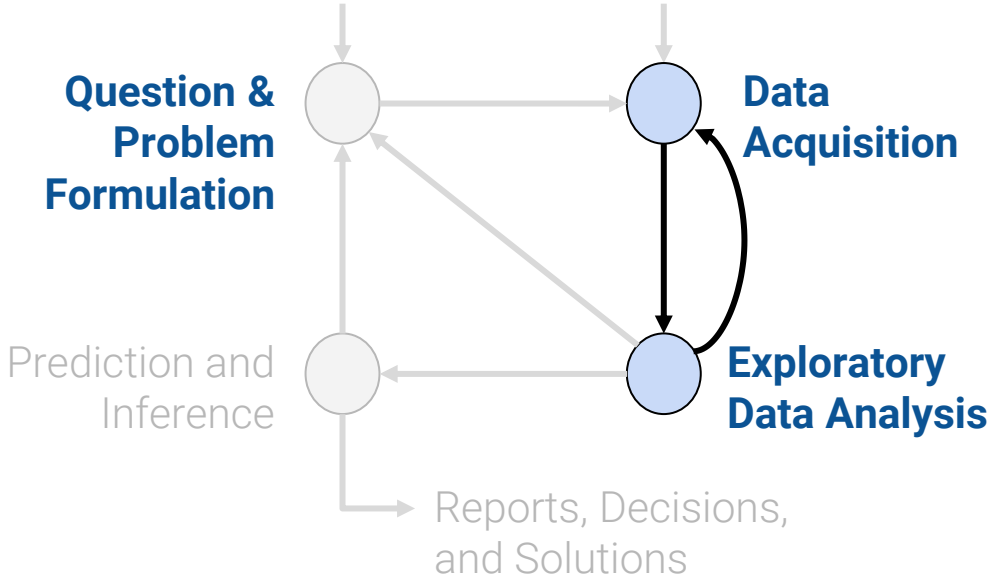
Variable Types

Multiple Files

More EDA/Wrangling

The Data Science Lifecycle is a Cycle

In practice, EDA informs whether you need more data to address your research question.



What is incidence?

Summary

What is already known about this topic?

The number of reported U.S. tuberculosis (TB) cases decreased sharply in 2020, possibly related to multiple factors associated with the COVID-19 pandemic.

What is added by this report?

Reported TB incidence (cases per 100,000 persons) increased 9.4%, from 2.2 during 2020 to 2.4 during 2021 but was lower than incidence during 2019 (2.7). Increases occurred among both U.S.-born and non–U.S.-born persons.

What are the implications for public health practice?

Factors contributing to changes in reported TB during 2020–2021 likely include an actual reduction in TB incidence as well as delayed or missed TB diagnoses. Timely evaluation and treatment of TB and latent tuberculosis infection remain critical to achieving U.S. TB elimination.

CDC Morbidity and Mortality Weekly Report (MMWR) 03/25/2022.

What is **incidence**? Why use it here?

How was "9.4% increase" computed?

Question: Can we **reproduce** these rates using government data?

CDC: **source**, 2021

Defining incidence

From the CDC report: **TB incidence** is computed as "Cases per 100,000 persons using mid-year population estimates from the U.S. Census Bureau."

- Incidence is useful when comparing case rates across differently sized populations.

$$\textbf{TB incidence} = \frac{\text{\# TB cases in population}}{\text{\# groups in population}}$$

(group: 100,000 people)

$$= \frac{\text{\# TB cases}}{(\text{population}/100{,}000)}$$

$$= \frac{\text{\# TB cases}}{\text{population}} \times 100{,}000$$

We don't have U.S. Census population data in our DataFrame.
We need to acquire it to verify incidence!

Demo Slides

Structure: Primary Keys and Foreign Keys

Sometimes your data comes in multiple files:

- Often data will reference other pieces of data.
- Alternatively, you will collect multiple pieces of related data.

Use `pd.merge` to **join** data on **keys**.

Customers.csv

CustID	Addr
171345	Harmon..
281139	Main ..

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

Purchases.csv

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

Structure: Primary Keys and Foreign Keys

Sometimes your data comes in multiple files:

- Often data will reference other pieces of data.
- Alternatively, you will collect multiple pieces of related data.

Use `pd.merge` to join data on **keys**.

Primary key: the column or set of columns in a table that determine the values of the remaining columns

- Primary keys are unique, but could be tuples.
- Examples: SSN, ProductIDs, …

Primary Key

Customers.csv

CustID	Addr
171345	Harmon..
281139	Main ..

Primary Key

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

Primary Key

Purchases.csv

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

Structure: Primary Keys and Foreign Keys

Sometimes your data comes in multiple files:

- Often data will reference other pieces of data.
- Alternatively, you will collect multiple pieces of related data.

Use `pd.merge` to join data on **keys**.

Primary key: the column or set of columns in a table that determine the values of the remaining columns

- Primary keys are unique, but could be tuples.
- Examples: SSN, ProductIDs, …

Foreign keys: the column or sets of columns that reference primary keys in other tables.

More later when we see SQL.
Stay tuned!

Primary Key

Customers.csv

CustID	Addr
171345	Harmon..
281139	Main ..

Foreign Key

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

Purchases.csv

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

More EDA/Wrangling

Exploring Tabular Data

Record Granularity

Variable Types

Multiple Files

More EDA/Wrangling

What else?

More next time

Structure -- the “shape” of a data file

Granularity -- how fine/coarse is each datum

Scope -- how (in)complete is the data

Temporality -- how is the data situated in time

Faithfulness -- how well does the data capture “reality”

What else?

Structure -- the “shape” of a data file

Granularity -- how fine/coarse is each datum

Scope -- how (in)complete is the data

Temporality -- how is the data situated in time

Faithfulness -- how well does the data capture “reality”

Faithfulness: Do I trust this data?

Does my data contain **unrealistic or "incorrect" values**?

- Dates in the future for events in the past
- Locations that don't exist
- Negative counts
- Misspellings of names
- Large outliers

Does my data violate **obvious dependencies**?

- E.g., age and birthday don't match

You will explore this more in homework. Stay tuned!

Was the data **entered by hand**?

- Spelling errors, fields shifted …
- Did the form require all fields or provide default values?

Are there obvious signs of **data falsification**?

- Repeated names, fake looking email addresses, repeated use of uncommon names or fields.

LECTURE 4

Data Wrangling and EDA, Part I

Content credit: Narges Norouzi, Lisa Yan, Josh Hug