Exploratory Data Analysis

Data Wrangling and Exploratory Data Analysis: An Infinite Loop

Key Data Properties to Consider in EDA

- Structure
  - File format
  - Variable types
  - Primary and Foreign Keys
- Granularity, Scope, Temporality
- Faithfulness (and Missing Values)

EDA Demo: Mauna Loa CO2

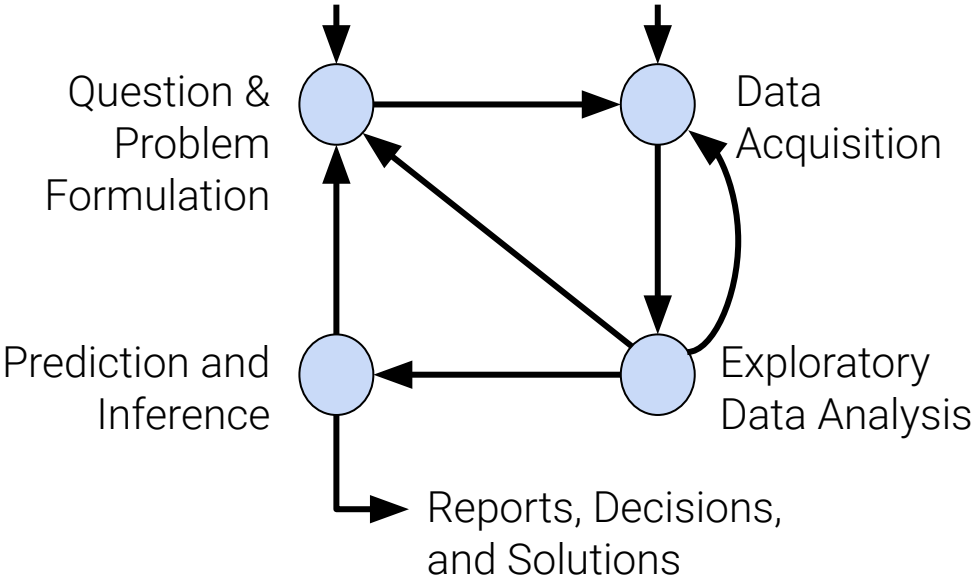
Now

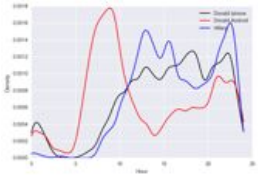
Congratulations!!!

You **have collected** or **<u>have been given</u>** a box of data.

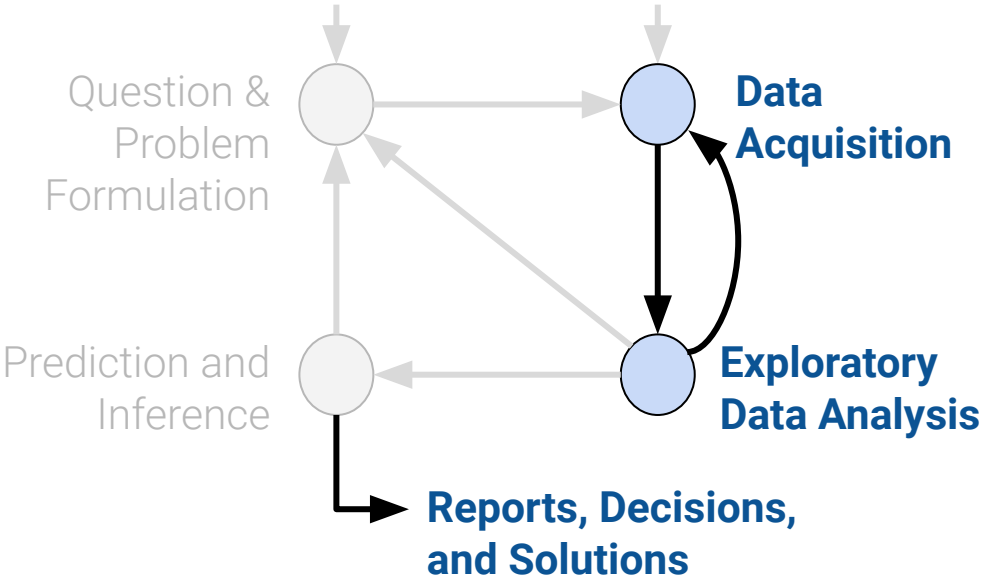
What do you do next?

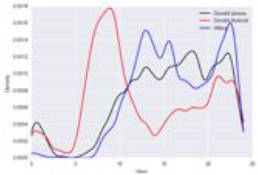
Plan for next few lectures





Plan for next few lectures





(today)

Data Wrangling
Intro to EDA

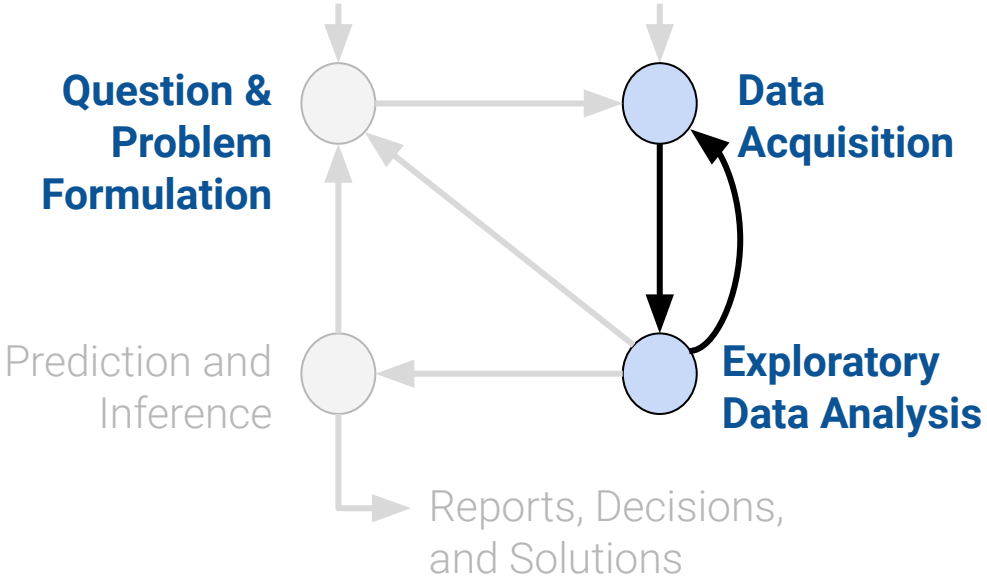
(next class)

Plots and variables
Seaborn

This week (Processing & Visualizing Data)

The Data Science Lifecycle is a Cycle

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



Key Data Properties to Consider in EDA

Structure -- the “shape” of a data file

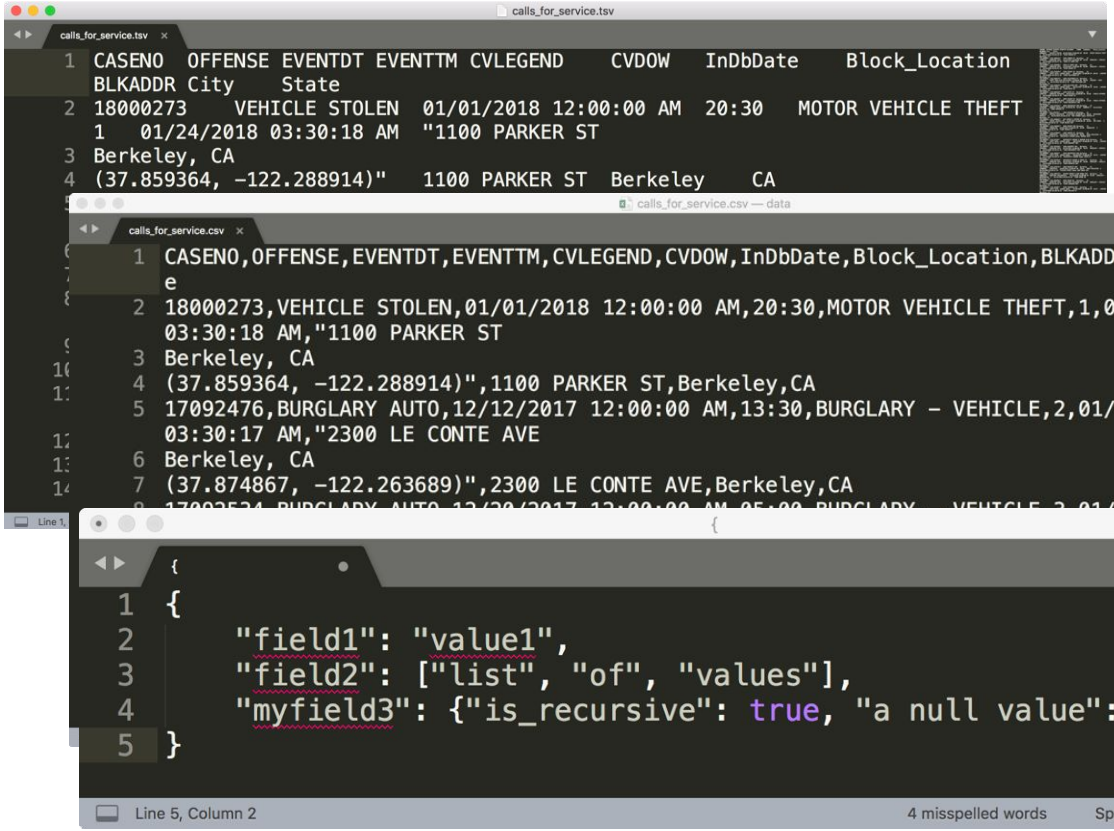
Granularity -- how fine/coarse is each datum

Scope -- how (in)complete is the data

Temporality -- how is the data situated in time

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

How are these data files formatted?



TSV

Tab separated values

CSV

Comma separated values

JSON

Which is the best? It depends on your use case.

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.

The U.S. Jurisdiction **variable**

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

It has two common properties:

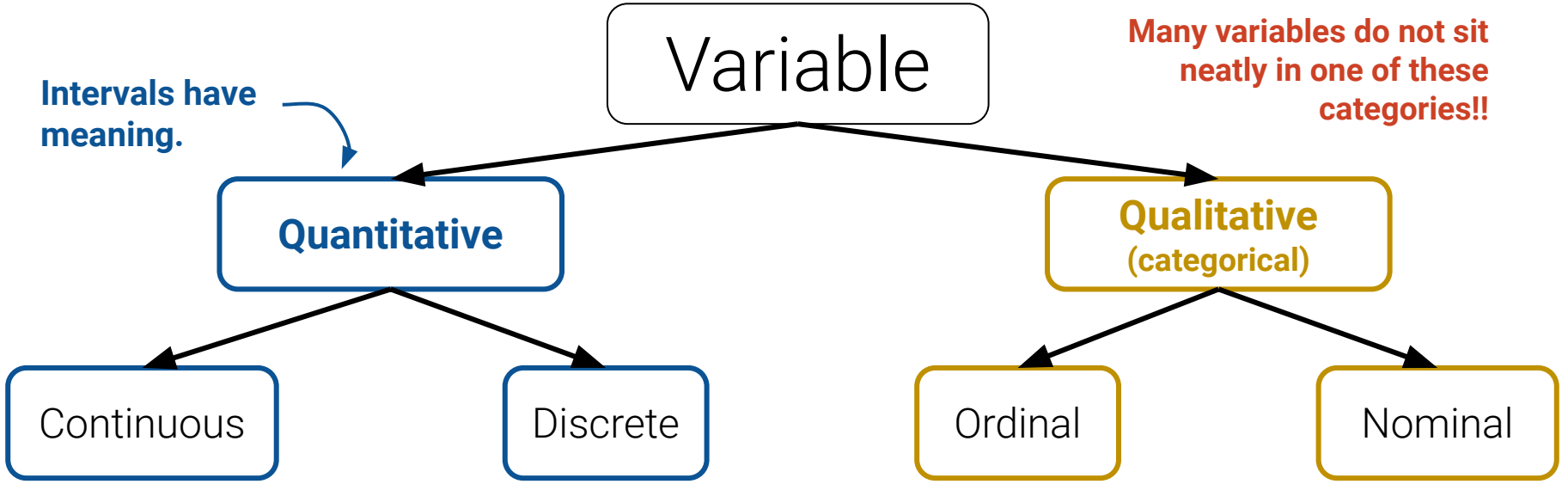
- **Datatype/Storage type**:
  How each variable value is stored in memory. `df[colname].dtype`
  - integer, floating point, boolean, object (string-like), etc.

  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
- CU ID number

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

Are the data in a standard format or encoding?

- Tabular data: CSV, TSV, Excel, SQL
- Nested data: JSON or XML

Are the data organized in **records** or nested?

- Can we define records by parsing the data?
- Can we reasonably un-nest the data?

Does the data reference other data?

- Can we join/merge the data?
- Do we need to?

What are the **fields** in each record?

- How are they encoded? (e.g., strings, numbers, binary, dates …)
- What is the type of the data?

Structure -- the “shape” of a data file

Granularity -- how fine/coarse is each datum

Scope -- how (in)complete is the data

Summary

You will do the most data wrangling when analyzing the structure of your data.

Demo Slides

What is the following file format?

Mauna Loa Observatory CO2 levels (NOAA)

How do we load these data into Pandas?

`pd.read_csv`? `pd.DataFrame`?

Often files will have mixed file formats, incorrect extensions or no extension at all.
You may need to explore the actual raw data file!

Demo Slides

What are our Variable Feature Types?

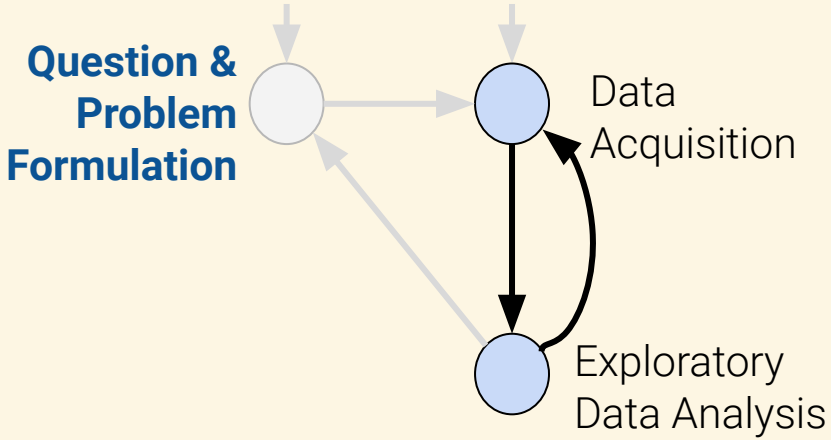
EDA step:
Understand what each record, each feature represents

First, **read file description**:

- All measurement variables (**`average`**, **`interpolated`**, **`trend`**) are monthly mean CO2 monthly mean mole fraction
  - i.e. monthly average CO2 ppm (parts per million)
  - Computed from daily means
- **`#days`**: Number of daily means in a month (i.e., # days equipment worked)

What variables define the first three columns?

- Year, month, and date in decimal



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

What does each **record** represent?

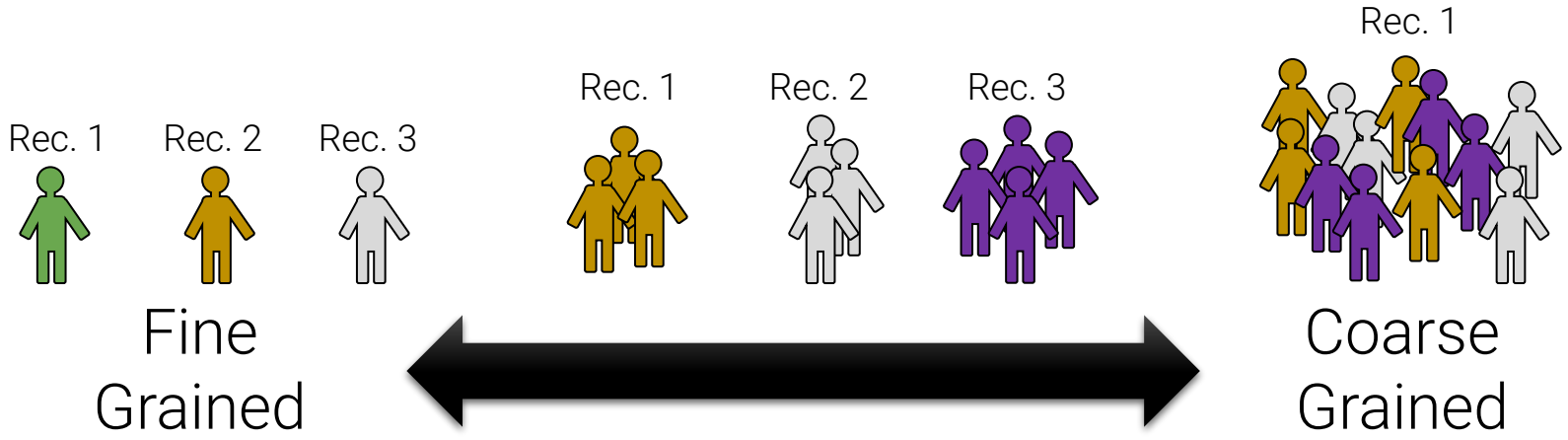
- Examples: a purchase, a person, a group of users

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, …



Scope

Does my data cover my area of interest?

- **Example**: I am interested in studying crime in Colorado but I only have Boulder crime data.

Are my data too expansive?

- **Example**: I am interested in student grades for CSCI 3022 but have student grades for all statistics classes.
- **Solution: Filtering** ⇒ Implications on sample?
    - If the data is a sample I may have poor coverage after filtering …

Does my data cover the right time frame?

- More on this in Temporality…

Scope

Does my data cover my area of interest?

- **Example**: I am interested in studying crime in Colorado but I only have Boulder crime data.

Are my data too expansive?

- **Example**: I am interested in student grades for CSCI 3022 but have student grades for all statistics classes.
- **Solution: Filtering** ⇒ Implications on sample?
  - If the data is a sample I may have poor coverage after filtering …

Does my data cover the right time frame?

- More on this in Temporality…

The **sampling frame** is the population from which the data were sampled. Note that this may not be the population of interest.

How complete/incomplete is the frame (and its data)?

- How is the frame/data situated in place?
- How well does the frame/data capture reality?
- How is the frame/data situated in time?

Temporality

Data changes – when was the data collected/last updated?

Periodicity — Is there periodicity? Diurnal (24-hr) patterns?

What is the meaning of the time and date fields? A few options:

- When the “event” happened?
- When the data was collected or was entered into the system?
- Date the data was copied into a database? (look for many matching timestamps)

Time depends on where! (**time zones** & daylight savings)

- Learn to use `datetime` python library and Pandas `dt` accessors
- Regions have different datestring representations: 07/08/09?

Are there strange null values?

- E.g., **January 1st 1970**, January 1st 1900…?

Temporality: Unix Time / POSIX Time

Time measured in seconds since **January 1st 1970**

- Minus leap seconds …

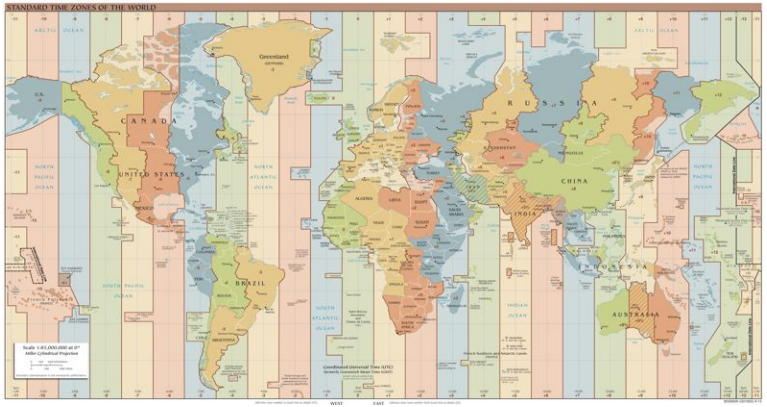
Unix time follows Coordinated Universal Time (UTC)

- International time standard
- Measured at 0 degrees latitude
  - Similar to Greenwich Mean Time (GMT)
- No daylight savings
- Time codes

Time Zones:

- San Francisco (UTC-8) without daylight savings

Feb 1, 2022 3:00pm Pacific
1643756400



https://en.wikipedia.org/wiki/Coordinated_Universal_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”

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

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.

Signs that your data may not be faithful (and proposed solutions)

Truncated data
Early Microsoft Excel limits: 65536 Rows, 255 Columns

Duplicated Records or Fields

Identify and eliminate (use primary key).

Spelling Errors

Apply corrections or drop records not in a dictionary

Units not specified or consistent

Infer units, check values are in reasonable ranges for data

Time Zone Inconsistencies

Convert to a common timezone (e.g., UTC)

- Be aware of consequences in analysis when using data with inconsistencies.
- Understand the potential implications for how data were collected.

Missing Data???

Examples

" "	1970, 1900
0, -1	NaN
999, 12345	Null

NaN: "Not a Number"

Missing Data/Default Values: Solutions

A. Drop records with missing values

- Probably most common
- **Caution**: check for biases induced by dropped values
  - Missing or corrupt records might be related to something of interest

B. Keep as NaN

C. Imputation/Interpolation: Inferring missing values

- **Average Imputation**: replace with an average value
  - Which average? Often use closest related subgroup mean.
- **Hot deck imputation**: replace with a random value
- **Regression imputation**: replace with a predicted value, using some model
- **Multiple imputation**: replace with multiple random values.

Examples

```
" "
0, -1
999, 12345
1970, 1900
NaN
Null
```

Missing Data/Default Values: Solutions

A. Drop records with missing values

- Probably most common
- **Caution**: check for biases induced by dropped values
  - Missing or corrupt records might be related to something of interest

B. Keep as NaN

C. Imputation/Interpolation: Inferring missing values

- **Average Imputation**: replace with an average value
  - Which average? Often use closest related subgroup mean.
- **Hot deck imputation**: replace with a random value
- **Regression imputation**: replace with a predicted value, using some model
- **Multiple imputation**: replace with multiple random values.

} (beyond this course)

Choice affects bias and uncertainty quantification (large statistics literature)

Essential question: why are the records missing?

Demo Slides

The Search for the Missing Values

EDA step:
Hypothesize why these values were missing, then use that knowledge to decide whether to drop or impute missing values

From file description:

- **`-99.99`**: missing monthly average **`Avg`**
- **`-1`**: missing value for **`# days`** that the equipment was in operation that month.

Which approach?

- Drop missing values
- Keep missing values as NaN
- Impute

Summary: Dealing with Missing Values

Mauna Loa Observatory CO2 levels (NOAA)

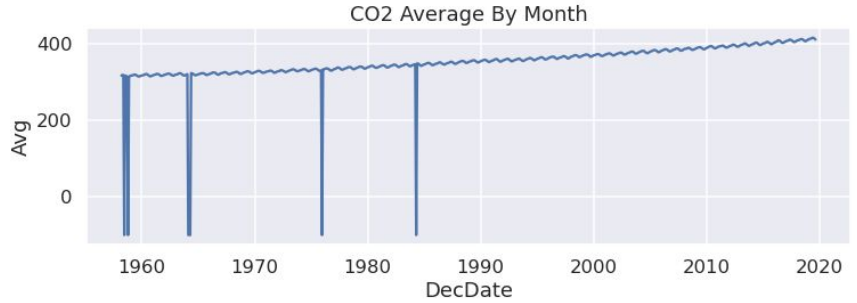
`-99.99`: missing monthly average **`Avg`**

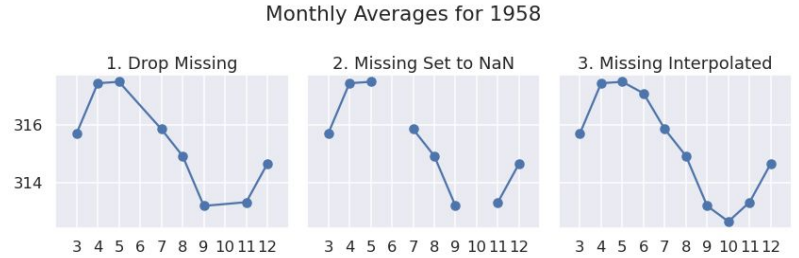
Option A: Drop records

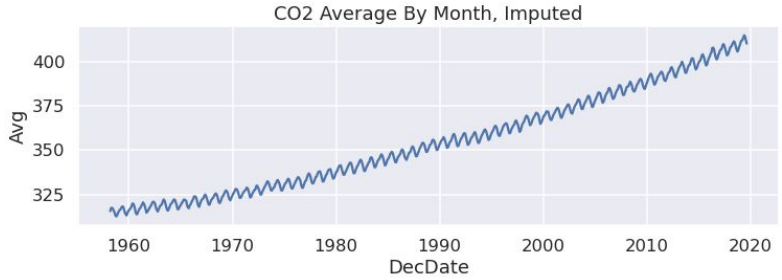
Option B: NaN missing values

Option C: **Impute** using interpolated column **`Int`**

All 3 are probably fine since few missing values, but we chose Option 3 based on our EDA.







With **numeric data**, you generally wrangle as you do EDA.

With **text data**, **wrangling is upfront** and requires new tools: **Python string manipulation** and **regular expressions**.

Summary: How do you do EDA/Data Wrangling?

Examine **data and metadata**:

- What is the date, size, organization, and structure of the data?

Examine each **field/attribute/dimension** individually

Examine **pairs of related dimensions**

- Stratifying earlier analysis: break down grades by major …

Along the way:

- **Visualize**/summarize the data
- **Validate assumptions** about data and collection process. Pay particular attention to when data were collected.
- Identify and **address anomalies**
- Apply data transformations and corrections (next lecture)
- **Record everything you do**! (why?)
    - Developing in Jupyter Notebooks promotes reproducibility of your own work.

Aside: An update to the Mauna Loa Dataset

https://gml.noaa.gov/ccgg/trends/data.html

Due to the eruption of the Mauna Loa Volcano, measurements from Mauna Loa Observatory were suspended as of Nov. 29, 2022. Observations from December 2022 to July 4, 2023 are from a site at the Maunakea Observatories, approximately 21 miles north of the Mauna Loa Observatory. Mauna Loa observations resumed in July 2023.

NPS

LECTURE 6

Data Wrangling and EDA