

Lecture 8

Data Cleaning 2





Last Class Notebook File

https://colab.research.google.com/drive/10kS9JcQDsvwTMhX_P5Pj3fnsHQQ6Wrwf?usp=sharing

RECAP: What is Data Cleaning

Data cleaning is the process that removes data that does not belong in your dataset. It does this by:

- fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset.



Recap: Missing Data



Missing data can be categorized as:

1. Data Missing Completely at Random (MCAR)
2. Data missing at random (MAR)
3. Data missing not at random (MNAR)



1. MCAR: Data missing Completely at random

There is no pattern to what data is missing

- In practical terms, for all possible values in that column, there is an equal chance of that value being missing
- The missing data points are randomly distributed and have no relationship with any values, whether observed (successfully collected) or unobserved (missing).



Characteristics of MCAR Data

- **Independence from Data:** The probability of any data point being missing is unrelated to the values of any variables in the dataset.
- **No Bias Introduced:** Since the missingness is random, it does not introduce systematic bias into the analysis.
- **Analysis:** Standard statistical techniques remain valid since the missing data do not affect the dataset's overall characteristics.

Example: MCAR



"Imagine a big test where each student gets different random questions. Because the questions are given out randomly, we can assume all students are similar in their abilities. Even though some questions are missing for each student, we know this missing data is because of the random way the questions are given."

Notice: The missing data are entirely unrelated to any other data or observations.

- In the test example, each student gets a random subset of questions.
- So, the missing questions are randomly distributed and not influenced by any student characteristics or other observed data.



2. MAR: Data missing at random

The missingness is related to some of the **observed data** but not the missing data itself

- Missing data are systematically related to other observed variables, but not directly to the missing data itself

Example: MAR



"Imagine a pop quiz given to all students on a single day. We have scores for students who were present, but scores are missing for those who were absent. When we check the attendance records, we notice that students with **poor attendance tend to have lower quiz scores** and more missing data."

Notice: The missing data in this scenario are related to an observed variable (attendance record) but not directly to the missing quiz scores.

- if we assume that being absent on quiz day was random once we consider their attendance history, we can use the scores we have to make educated guesses about the missing scores.

HOW TO HANDLE

Data Missing at Random/ Completely Random



Data Missing at Random

For data that is categorical, it's fine! "Missing" can just become a new category.

Ex: If our category is major, we could have { "Computer Science", "Philosophy", "Did not answer" }

For numerical data, most algorithms we use **cannot accept NaN**. So what do we do?



Solution 1: Drop it

If 1% of your data has missing rows and you have a terabyte of it, just drop all rows that have missing stuff.

```
df.dropna()
```

Listwise Deletion: Remove any rows with missing data. This works well when the proportion of missing data is small.

Pairwise Deletion: Use only the data points where both variables being analyzed (e.g., income and education) have non-missing values, excluding cases with missing values for either variable from that specific analysis.

Solution 2: Apply Imputation!



Imputation is the process of replacing missing values with estimates (new values).

Data may be imputed in several ways:

- **Mean/ Median/Mode Imputation:** Imputation Using (Mean/Median/Mode) Values
- **Hot Imputation:** the value is selected from other instances in the same dataset.
- **Cold Imputation:** the value is selected from other instances in the different dataset ([use external data](#)).
- **Multiple imputation**
- etc.

CHECK LECTURE 7 for Mean/ Median/Mode Imputation

Hot-Deck Imputation



Hot-deck imputation: is a method used to fill in missing values in a dataset by replacing them with **values from similar record/row**.

How it works? Find a **row that is most similar** to the one with the missing value, and copy that value. We can also average over several similar data-points.

- If someone is missing a grade for CMSC 320, find a student who has taken all the same classes as them and copy over their CMSC 320 grades.

Example: Hot-Deck Imputation

Hot-deck imputation Example: Suppose you have a dataset containing information about the ages of individuals, but some ages are missing

ID	Gender	Age
101	Male	32
102	Female	NaN
103	Male	NaN
104	Female	45
105	Male	NaN
106	Female	28

Example: Hot-Deck Imputation

Identify Similar Cases: determine which cases (rows) are similar (such as gender or other relevant characteristics) or close to the one with missing data.

Impute with Similar Values: For each missing age value, find a similar individual (a "donor") from the dataset. In this example, let's use the same gender as a criterion for similarity.

ID	Gender	Age
101	Male	32
102	Female	NaN
103	Male	NaN
104	Female	45
105	Male	NaN
106	Female	28

Example: Hot-Deck Imputation

Hot-deck imputation: After applying hot-deck imputation, your table might look like this



ID	Gender	Age
101	Male	32
102	Female	28
103	Male	32
104	Female	45
105	Male	32
106	Female	28



Bayesian Imputation

A statistical technique used for imputing missing data in a dataset using Bayesian methods.

- **Bayesian statistics is based on Bayes' theorem**, which allows for the estimation of unknown parameters by combining prior information (prior beliefs) and observed data.
- **Multiple Possibilities:** Unlike simpler imputation methods that rely on point estimates, Bayesian imputation, **it thinks about many values (a range of possible values) for missing data, not just one guess (single estimate point).**



Bayesian Imputation

For categorical data, for example:

$$p(\text{possible_value} \mid \text{features}) = \frac{p(\text{features} \mid \text{possible_value}) * p(\text{possible_value})}{p(\text{features})}$$

Evaluate each of these for each feature and assign!



Bayesian Imputation: another example

Suppose you have a dataset about fruits with two features: color and shape. You want to impute the missing values for the "Type of Fruit" category.

Color	Shape	Type of Fruit
Red	Round	Apple
Yellow	Long	Banana
Green	Round	NaN
Yellow	Round	NaN
Red	Round	NaN

You want to fill in the missing 'Type of Fruit' values using Bayesian imputation, which involves estimating the probabilities of each fruit type based on the observed data.



Multiple Imputation (common technique for MAR)

Handle missing data by creating multiple plausible values for each missing data point. These multiple imputed datasets are analyzed separately, and then combined to obtain statistically valid results that reflect the uncertainty due to missing data.

- Example: Use Age and Gender to predict and fill in missing Income values multiple times, then average the results.



Multiple Imputation: Example

Imagine you're surveying people about their height and weight, but some participants don't answer one or both questions, leaving gaps in your data.

- **Multiple Imputation:** Instead of guessing one value to fill each gap, you make several copies of the dataset. Each copy uses statistical methods (like regression) to predict the missing values with plausible ones.
 - For example, if weight is missing for some taller people, you use their height and other available data (like gender) to predict a range of possible weights in each copy.
- **Final Analysis:** The final results consider all these predictions, showing the range of possible outcomes for each missing value.

HOW TO HANDLE

Data Missing Not at Random (MNAR)

3. MNAR: Data not missing at random

The worst: also known as Non-ignorable (NI) missing data or informative missing data or NMAR.

When data is not missing at random, it means that the likelihood of data being missing **depends on the unobserved data (missing data) itself**, even after considering the observed data.

- In simpler terms, the probability of data being missing is **related to the values of the missing data themselves, rather than just the observed data**.

This is the most challenging type of missing data to handle, as the missingness is related to the unobserved (missing data) data itself!



Example 1: MNAR

Data not missing at random is subject to some sort of systematic bias.

Example: Suppose we have a dataset with information about **individuals' income levels (observed or collected data)**, and some individuals choose **not to report their income (missing data or unobserved data)**. If this unobserved data includes individuals with higher incomes who are less likely to report their income, it introduces a systematic bias into the dataset

Here, Missingness mechanism is not random; rather, it's influenced by the values of the variable being measured (income).



Example 3: MNAR

“Imagine a survey that asks students about their GPA. Some students with lower GPAs might choose not to answer the survey.”

The missingness of the GPA data (students who did not answer the survey) is directly related to their unobserved characteristic (lower GPA)



Biases: Data Not Missing at Random

Data not missing at random is subject to some sort of systematic bias

A. Missing rows in series

- data might be missing in consecutive rows or periods. Example, missing data might occur every weekend or during specific months

B. Missing based on value

- in a survey about income, people with higher incomes might be less likely to disclose their exact income

C. Boundary Conditions

- missing when it reaches certain boundary conditions or thresholds. Example, in a temperature dataset, values might be missing when they fall below a certain minimum temperature or exceed a maximum temperature



A. Missing Data in Series

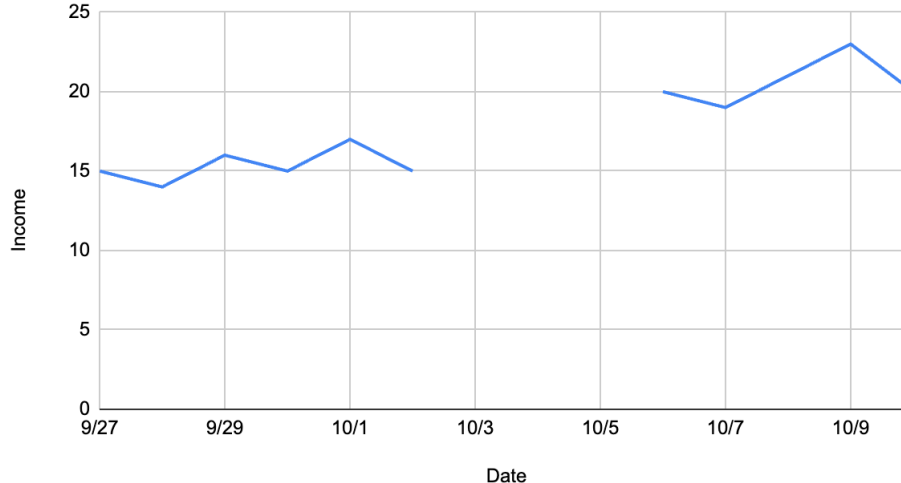
More Examples: Sometimes, you'll be missing data in a series:

- The middle 50 entries in a survey got dropped
- You lost three days of ad-spend info
- A single sensor turned off when measuring water quality at multiple points in a river

Imputation Using Framing Data: Approach for handling missing data

Imputation using framing data means filling in those missing values based on similar information from other observations.

Income vs. Date



We can use information from similar time periods or frames to estimate or impute missing income values over time.

E.g., Look at income patterns from similar months in the past

B. Data Missing Based on Value

Data is missing based on the value or characteristics of the variable itself.

This is very, very bad.

Not a lot of easy ways to fix this! Especially if you can't determine how it interacts with data that is also missing at random.

For example:

- People who vote for Trump are 30% less likely to say who they're voting for on a survey. (*related to a specific value (support for Trump)*)
- Your tractor telemetry cuts out at random times, but *also* has a 50% chance of cutting out if your tractor is stuck in the mud (*related to a specific value (tractor being stuck in the mud)*)





Try

- **Sensitivity Analysis:** explores how your results might change based on different assumptions about the missing data.
 - Assume different scenarios for high and low missing incomes and observe how these assumptions affect the overall analysis.
- **Pattern Mixture Models:** create separate models for groups with and without missing data to understand the impact of missingness.
 - Create different models for groups with and without missing data to understand the impact of missingness.



C. Dealing With Boundary Conditions

Depends on the circumstances.

- I. **Drop everything on the boundary** and only work with the stuff within (if you want to just do predictions within the boundary)
- II. **Extrapolate** the same distribution outside the boundary
- III. **Get more data**

Example: Dealing With Boundary Conditions



Example: Temperature Sensor Data: Imagine you have a dataset that **records outdoor temperatures (in degrees Celsius) over time**. However, the sensor used to collect data has a known boundary condition—it can't measure temperatures below -10°C or above 40°C accurately.

- I. **Exclude Data at the Boundary:** Remove data points that **fall outside** the valid temperature range (-10°C to 40°C) if your analysis focuses on typical weather conditions within that range.

Example: Dealing With Boundary Conditions



Example: Temperature Sensor Data: Imagine you have a dataset that **records outdoor temperatures (in degrees Celsius) over time**. However, the sensor used to collect data has a known boundary condition—it can't measure temperatures below -10°C or above 40°C accurately.

- **II. Extrapolate Data:** Use statistical methods to estimate temperatures below -10°C or above 40°C if you want to predict weather conditions beyond what your sensor can measure.

Note: Extrapolation means **using existing data to make predictions or estimates for values** that are outside the range of the data you have.

Example: Dealing With Boundary Conditions



Example: Temperature Sensor Data: Imagine you have a dataset that **records outdoor temperatures (in degrees Celsius) over time**. However, the sensor used to collect data has a known boundary condition—it can't measure temperatures below -10°C or above 40°C accurately.

- **III. Acquire More Data:** If your analysis requires accurate temperature readings both below -10°C and above 40°C , then **acquire more data from a sensor that can operate in those extreme conditions**. More data points within the boundary can provide a better understanding of temperature variations.

Incorrect Data



Detecting Incorrect Data

What does it mean for data to be incorrect?



Types of Incorrect Data

- People lied to you
- Your instrument broke
- You've been recording the wrong metrics
- You have two identical entries with different values
- Illegal values
- Unclear default values

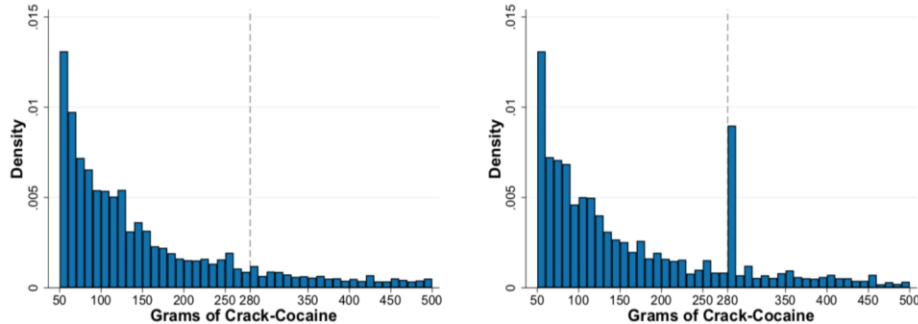
Detecting Incorrect Data



Detecting incorrect data involves **looking for anomalies or patterns that deviate** from what's expected or reasonable in your dataset.

Detecting Incorrect Data

Figure 1. Changing Distribution of Drug Amounts Around 280g Pre- and Post-2010.
(a) 1999-2010 (b) 2011-2015



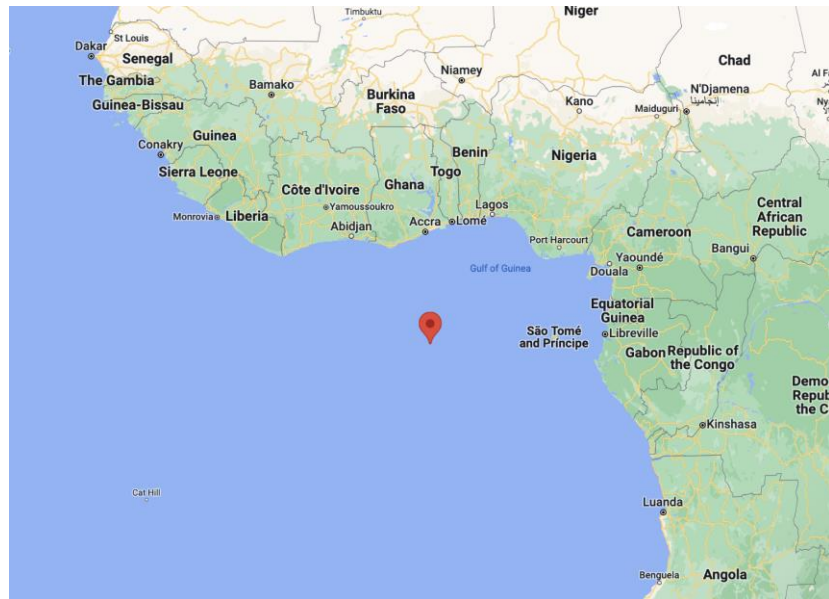
How do we find data that's incorrect?

- **Attractors:** Check for unusual spikes or concentrations of data points in specific categories or values.
 - Imagine if there were very few people who were 5'5", but a big spike in people who were six feet
- **Discontinuities:** Look for abrupt changes/ shift or discontinuities in your data.
 - Once the limit for how much weed constituted a felony changed, there was a significant discontinuity centered around that threshold (People might adjust their behavior based on the new legal limit).
 - This helps us understand how the law affected the situation or data

Detecting Incorrect Data

How do we find data that's incorrect?

- Modes that don't make sense: Identify **modes (frequent values) in your data that don't align** with the expected patterns.
 - While analyzing geographical data, finding many people living in latitude/longitude 0,0



Detecting Incorrect Data: Boundary Conditions

How do we find data that's incorrect?

- **Data outside valid bounds:** Check for data points that **fall well outside the valid range or bounds of a variable.**
 - People who have been playing video games for one million hours
- Suddenly a field increases an order of magnitude for no reason



Boundary Conditions: Why can occur?



Boundary conditions occur when your ability to take in data hits some sort of artificial cap (limits or restrictions on the data you can collect or record)

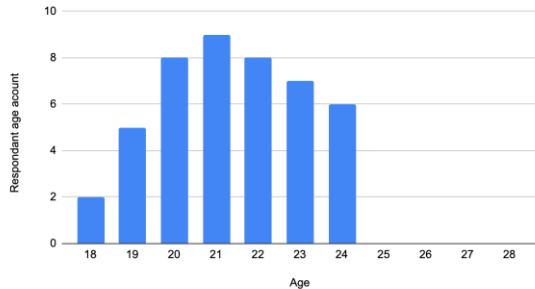
- **Instrument Limits:** your instruments have a maximum measurement: For example, a **thermometer might not be able to measure temperatures higher than 100 degrees.**
- **Database Constraints:** your database caps certain measures: **it might not allow values larger than a particular number.**
- **Survey Limits:** Your survey values only go up to a certain amount: If you're conducting a survey, there **may be a maximum value or range for the responses you collect.** For instance, if you ask about income, your survey might only have options up to a certain income level.

Signs of Boundary Conditions

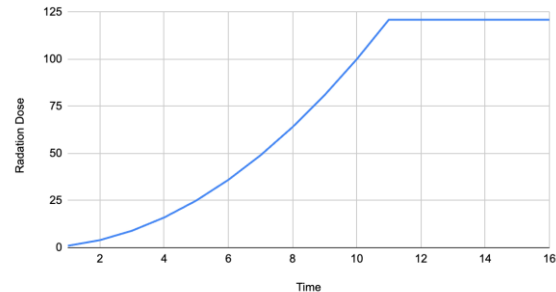
You will see:

- Your data has a discontinuity at a certain value, after which there is nothing

Respondant age account vs. Age



Radation Dose vs. Time





Instrument Error

Refers to problems or **mistakes that can happen with measurement instruments or devices**



Instrument Error: Examples

Some detection apparatus begins to function abnormally.

- **A scale is incorrectly tared:** Imagine you have a weighing scale, **but it's not set to zero correctly**; shows the wrong weight because it didn't start from zero.
- **A sensor begins to lose sensitivity and so reports lower readings before being replaced:** a temperature sensor might become **less sensitive, so it starts showing lower temperatures** than what's actually there. It might work fine again after it's fixed or replaced.
- **A microphone begins to pick up more noise:** This can make the recorded audio less accurate.



Instrument Error

Some detection apparatus begins to function abnormally.

- A scale is incorrectly tared
- A sensor begins to lose sensitivity and so reports lower readings before being replaced
- A microphone begins to pick up more noise

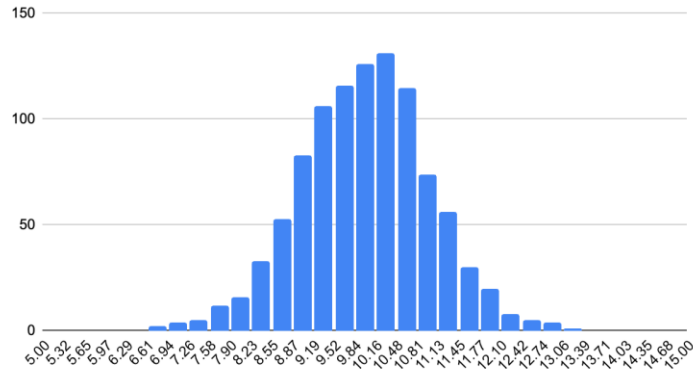
How to repair:

- **Review Past Data:** Look at past sensor data for that sensor or similar ones, and
- Use the insights gained from past data **adjust the distribution** to be in line with the old one
- This only works if you understand what sort of error you have!

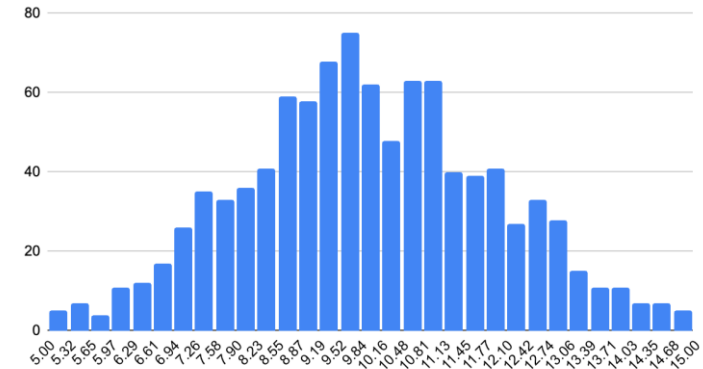
Instrument Error

We want to transform the malfunctioning sensor to have the same properties as the typical sensor (no error)

Typical Sensor

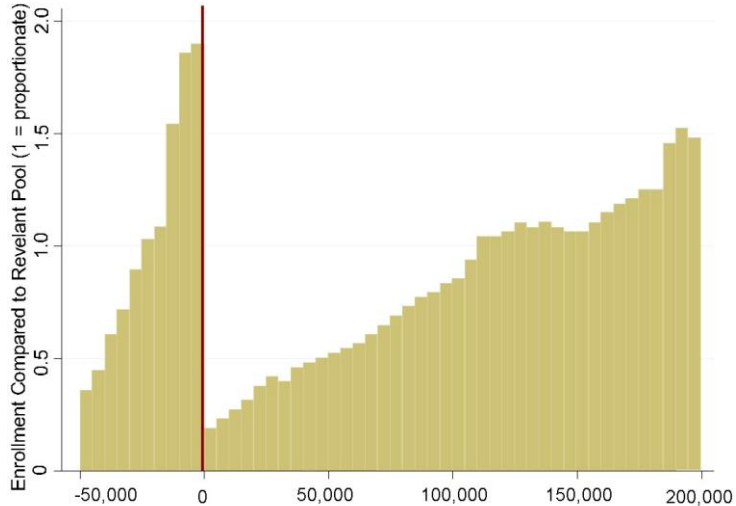


Malfunctioning Sensor

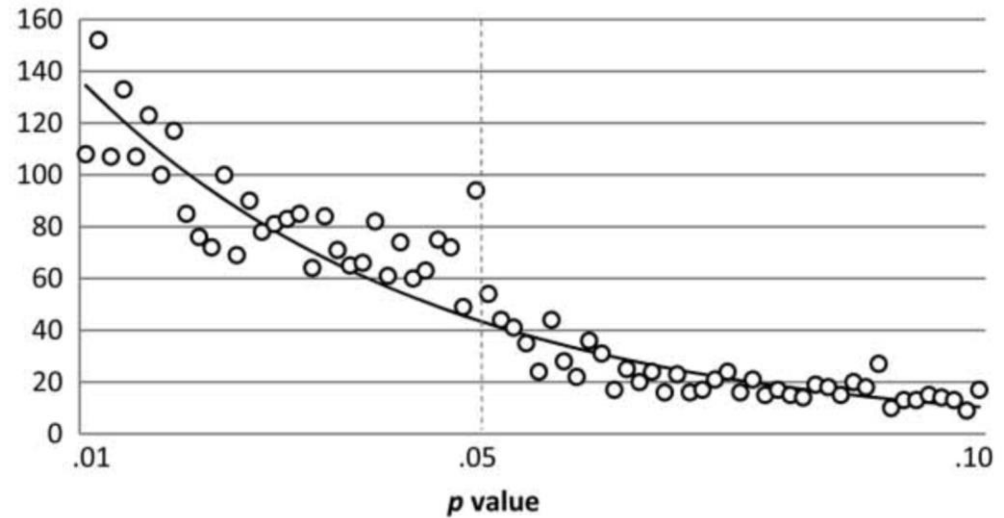


This adjustment might involve **shifting values, scaling them, or applying other corrective measures.**

Great Examples of Lies



Family's AGI versus the AGI Student Needs to Qualify for a Pell Grant, in \$5000 Bins
e.g. in the bin just below 0, family could earn between \$0 and \$5000 more and still qualify for Pell



SUMMARY



- **Messy Data:** Data can be messy due to missing values, column combinations, or joining multiple tables.
- **Data Cleaning:** The first step in data analysis is cleaning the data to ensure it's accurate and ready for analysis.
- **Handling Obvious Issues:** Use functions like `apply()` for simple and obvious data issues, like correcting misspellings in categorical data.
- **Data Typing:** Ensure data is in the correct format; use `astype()` or `apply()` for type conversions.
- **Merging Tables:** Merge tables carefully, ensuring file formats match.
- **Evolving Labeling Schemes:** Adapt to changes in labeling schemes by dividing the dataset, inferring old values, or handling changes as needed.

SUMMARY



Dealing with Duplicates: Identify and remove exact or subtle duplicates to maintain data quality.

Outlier Detection: Use methods like z-scores to find and consider removing outliers.

Imputing Missing Data: Handle missing data using mean, median, mode imputation, or more advanced methods like Bayesian imputation.

Boundary Conditions: Be aware of data limits imposed by instruments, databases, or surveys and adjust analysis accordingly.

SUMMARY



Detecting Incorrect Data: Identify issues like attractors, discontinuities, strange modes, and data outside valid bounds.

Repairing Instrument Errors: Correct data from malfunctioning sensors based on past data or similar sensors.

Signs of Boundary Conditions: Look for sudden changes or gaps in data and observe patterns in visualizations.

Data cleaning is a crucial step in data analysis to ensure the accuracy and reliability of your results.



Additional Reading Slides



Bayesian Imputation

A statistical technique used for imputing missing data in a dataset using Bayesian methods.

- **Bayesian statistics is based on Bayes' theorem**, which allows for the estimation of unknown parameters by combining prior information (prior beliefs) and observed data.
- **Multiple Possibilities:** Unlike simpler imputation methods that rely on point estimates, Bayesian imputation, **it thinks about many values (a range of possible values) for missing data, not just one guess (single estimate point).**



Bayesian Imputation

For categorical data, for example:

$$p(\text{possible_value} \mid \text{features}) = \frac{p(\text{features} \mid \text{possible_value}) * p(\text{possible_value})}{p(\text{features})}$$

Evaluate each of these for each feature and assign!



Bayesian Imputation: Example

Let's say you have a dataset containing information about the ages of individuals, but some ages are missing. You want to impute (fill in) the missing ages using Bayesian imputation.

- Using Bayesian imputation, you **estimate the missing ages based on the observed data and the relationships between age, gender, and other relevant variables.**
- For instance, if most females in the dataset are in their 30s, and a female individual with a missing age has similar characteristics to those in their 30s, the algorithm may impute an age around 30 for that individual.



Bayesian Imputation: another example

Suppose you have a dataset about fruits with two features: color and shape. You want to impute the missing values for the "Type of Fruit" category.

Color	Shape	Type of Fruit
Red	Round	Apple
Yellow	Long	Banana
Green	Round	NaN
Yellow	Round	NaN
Red	Round	NaN

You want to fill in the missing 'Type of Fruit' values using Bayesian imputation, which involves estimating the probabilities of each fruit type based on the observed data.