Dataframes

Missing values from data

- Looking at my data, I realized that a person forgot to answer a question in a survey.

col_1	col_2
10	a
20	b
	c

- What to do?

```
data = {'col_1': [10, 20], 'col_2': ['a', 'b', 'c']}
df = pd.DataFrame(data)
# This throws an error
```

- Replace with a 0?
  - `data = {'col_1': [10, 20, 0], 'col_2': ['a', 'b', 'c']}`
  - `df['col_1'].mean()`
    - `np.float64(10.0)`
- One option is to replace with Not a Number (NaN)
  - `data = {'col_1': [10, 20, numpy.nan], 'col_2': ['a', 'b', 'c']}`
  - Need to import numpy
  - `df['col_1'].mean()` ignores NaN
    - `np.float(15.0)`

Data Wrangling

Data wrangling is made up of

- Cleaning the Data
- Transforming the Data
- Integrating the Data

Data Cleaning

- Fixing or removing incomplete, abnormal, or inconsistent data to enhance the quality for accurate analysis and storytelling
- Data that we might need to address could include, but not limited to
  - missing values
  - nonsensical values
  - different formats of the same value

    - duplicate instances

Transforming Data

- Involves organizing and formatting data into a structure that is useful for efficient analysis
- Example: One attribute contains the dates Sept. 16, 2025, but we may only care about the year. We would leave that attribute alone, but create a new attribute "Year" and extract the year from the original date.

Integrating Data

- Process of combining multiple datasets from one or more sources to maximize the amount of information from which to draw insight
- A few ways to think about how datasets relate to each other
    - They have the same set of attributes / variables, but different instances / observations
    - They could have the same instances / observations, but different attributes / variables
    - They could have some similar instances / observations and some similar attributes / variables, but not identical sets of either
- We will need to consider what will create a more informative final dataset for our goals

Cleaning the Data

Cleaning Missing Data

- Important to note that a 0 in data does not necessarily mean missing value
    - Example: Number of cats owned
        - 0 indicates we know that they don't own any cats
        - Missing value means that they could have cats, but we don't know
    - Replacing missing values with 0 can largely affect the analysis (e.g., the mean value)
- How to handle missing values
    1. Removing instances with missing values
    2. Estimate the value based on internal data
    3. Estimate the value based on of external data
    4. Replace with a value representing missing values

Handling Missing Values

1. Removing the entire instance

   - Simple
   - Removes other potentially useful information
   - If the missing value is for an attribute that you don't need to use, then you can leave the instance alone
   - Be careful when analyzing data
     - Example: mean of attribute A for all 100 people, but mean of attribute B for only 86/100 people because there were 14 missing values for attribute B. The means were taken over different instances. This is okay to do, but be transparent.

2. Estimating the value from internal data

   - Based on existing data from the dataset, we sometimes can estimate a value for the missing value
   - Example: Let's say we have 2 attributes (has cats / likes cats)
     - If someone forgot to answer a question on the survey about liking cats, but they stated that they have cats, then we can look at our dataset and find that most people that have cats do actually like cats.

object_id	type	material	depth	age
o1	Pottery	Clay	1	2000
o2	Jewelry	Gold	2	
o3	Tool	Bronze	1.5	3000
o4	Weapon	Iron	3.5	1500
o5	Coin	Silver	1.5	
o6	Pottery	Clay	2.5	2500
o7	Jewelry	Gold	1	
o8	Weapon	Bronze	3	2000
o9	Tool	Iron	4	
o10	Coin	Silver	1.2	1900
o11	Jewelry	Gold	2.7	
o12	Pottery	Clay	2.2	2400

Mean age $= (2000 + 300 + 1500 + 2500 + 2000 + 1900 + 2400)/7 = 2185.71$ (The new mean after putting this value in the place of the missing values, is the same as the mean before. We could also use the median)

object_id	type	material	depth	age
o1	Pottery	Clay	1	2000
o2	Jewelry	Gold	2	2185.71
o3	Tool	Bronze	1.5	3000
o4	Weapon	Iron	3.5	1500
o5	Coin	Silver	1.5	2185.71
o6	Pottery	Clay	2.5	2500
o7	Jewelry	Gold	1	2185.71
o8	Weapon	Bronze	3	2000
o9	Tool	Iron	4	2185.71
o10	Coin	Silver	1.2	1900
o11	Jewelry	Gold	2.7	2185.71
o12	Pottery	Clay	2.2	2400

3. Estimate the value from external data
    - We can look toward external data to help us determine what value to estimate for some missing value
    - Example: Estimating a person's weight based on other attributes such as height, from a larger dataset
4. Replace with a value representing missing values
    - Replace missing value with a new value reserved for missing values ("NA" / "missing" / "unknown" / "NaN")

Case Study – Criminology

- Missing data can cause serious problems in criminology
- They can obscure the types of crimes that have occurred, contribute to a lack of evidence or faulty evidence, and confuse criminal justice researchers
- Data is rarely collected without missing values and estimating the values have proven effective in criminology research
- The FBI's Uniform Crime Reporting (UCR) Program is a primary source of data on crime and incorporated multiple forms of estimating missing data into its data cleaning process

Cleaning Anomalous Values

- Anomalous Values – Values that deviate from what is expected or standard
- Types of anomalous values:
    1. Implausible Value
        - Value is not within an expected or possible range
    2. Extreme Value
        - Value is not noticeably different from others, potentially affecting the overall analysis in noticeable ways
    3. Incorrect Format
        - Data entry is in an unexpected format, requiring correction to match the anticipated data type
    4. Duplicate Record
        - Repeated data entries (instances) in a dataset

1. Implausible Value
    - Value is not within an expected range
    - Examples:
        - Height of a person is 20' 8"
        - Weight of a cat is -10 lbs

  - The time for a 100m race was 2.4s
- These values are clearly not possible
- We can replace them with missing values and then process them as missing values

2. Extreme Value

   - Value is noticeably different from others, potentially affecting the overall analysis in noticeable ways
   - These values can be accurate, but they are largely different than the remaining data points
   - Examples:
     - Heights of students in a class range from 5' 1" to 6' except one student who is 7' 2"
     - Annual earnings of a populace ranges from $45,000 to $90,000 except for one individual who has reported an annual earning of $1.5 million
   - If the value is extreme enough though, it can affect the analysis of the data, therefore, sometimes we may want to remove it from the analysis but be transparent when reporting
   - Can provide 2 different reports (with and without the extreme value)
   - Anomaly or outlier is an extreme data point so different from the majority that it may warrant correction or removal

3. Incorrect Formats

   - The data is in an unexpected format
   - Examples:
     - Age: Written as "twenty-two" but expected "22"
     - Favorite Color: Written as "bule" instead of "blue"
     - Favorite Pet: Written as "cats" instead of "cat"
   - Usually, we don't want to remove data if we don't have to
   - Therefore, for the above examples, it is easy to see what the value was supposed to be, or the format that we would want
     - We can automatically change mistakes to the format we expected through some simple code
   - Is there a way that we can help reduce the chance of incorrect formats during data collection?

4. Duplicate Records

   - An instance in the dataset has the exact same values, but does this indicate that it represents the duplication of an entity that is being represented by the same values, or do two different entities happen to have the same values

Height	Weight	Age
5' 8"	165 lbs	25
5' 8"	165 lbs	25

We cannot confidently that that the two instances above are the same entity.

ID	Height	Weight	Age
00002	5’ 8”	165 lbs	25
00002	5’ 8”	165 lbs	25

The two instances above are duplicates and we can remove 1

- Duplicate records can be created when merging datasets
- Unique identifiers are a great way to help us determine duplicate records