Data Types & Structures

- Take home: Understand the different types of data you may be asked to analyze and be familiar with the different formats they might come in

Why Data Types Matter

- What's the average of (High, Medium, and Low)?

Types of Variables

- Nominal/Categorical
- Ordinal
- Boolean
- Discrete
- Continuous
- Datetime

- These are important to know because it will affect how you validate, explore, and ultimately model the data
- For example, continuous data lends itself to regression, but what about nominal?

Nominal/Categorical

- Categories with no inherent order:
  - Gender, color, country
  - Quantify with counts, visualize with histograms

Ordinal

- Categories with meaningful order
  - Low/Medium/High, Levels of education
    - Matters statistically because rank order is important
  - Quantify with counts (can rank), histograms

Boolean/Binary

- Binary, logical values
  - T/F, Yes/No, 0/1
  - Quantify with counts, visualize with bar charts/histogram

Discrete/Count

- Countable, numeric values
  - # of orders, # of cars in a parking lot → typically whole numbers
  - Quantify with central tendency, variability, visualize with histograms

Continuous

- Measurable numeric values
  - Height, weight, dollars
  - Quantify with measures of central tendency, variability; visualize with histogram

Datetime

- Data that represents a point in time (can be in seconds, unix, yyyy-mm-dd, hh:mm:ss)
  - Quantify with timediffs (especially useful for sensor data with timestamps), Visualize with line plots

Data Formats and Structures

- Tables (CSVs, Excel, SQL)
  - Rows = records, Columns = variables
  - Schema example
  - Arrays and matrices (NumPy-style) → n-D
- Key-value (JSON, Python dicts)
  - Nested, flexible, but messy to model
- xTime series: a special case of tabular data
  - Often stored as a table with a timestamp index
  - Needs sorting, alignment, resampling

Examples

Example (CSV or SQL):

id	name	age	signup_date
1	Alice	30	2023-01-10
2	Bob	24	2023-02-15

→ Rows = records (people), Columns = variables

→ Schema: `id` (int), `name` (str), `age` (int), `signup_date` (datetime)

```
{
  "user": "alice",
  "age": 30,
  "preferences": {
    "theme": "dark",
    "notifications": true
  }
}
```

```
data = np.array([
    [1.2, 3.5, 5.1],
    [4.4, 0.8, 2.9]
])
```

timestamp	temperature
2023-01-01 00:00	21.5
2023-01-01 01:00	20.8
2023-01-01 02:00	19.9

Importing Data

- Almost invariably, data files will have mixed data types that you need to track carefully
- Don’t take averages of ID numbers or zip codes

Common Pitfalls

- Not paying attention to column headers and taking averages of IDs, zipcodes
- Not checking the sampling rate/time order of time series data

Work an Example

Tree ID	Height (m)	Width (m)	Color	Species	Estimated Age (yrs)	Alive/Dead	Number of Leaves	Lumber Value
1	21.5	1.92	Bright Green	Quercus macrocarpa	48	Alive	30200	High
2	19.6	1.78	Bright Green	Quercus rubra	71	Alive	27329	Medium
3	21.9	1.52	Bright Green	Quercus palustris	43	Alive	35332	Medium
4	24.6	1.31	Yellow-Green	Quercus robur	89	Alive	34621	Medium
5	19.3	1.71	Green	Quercus palustris	44	Alive	31334	Medium
6	19.3	1.62	Green	Quercus macrocarpa	59	Alive	38174	Medium
7	24.7	1.77	Olive Green	Quercus robur	114	Alive	46053	High
8	22.3	1.69	Green	Quercus palustris	76	Alive	20885	High
9	18.6	1.81	Green	Quercus macrocarpa	61	Alive	27734	Medium
10	21.6	1.34	Bright Green	Quercus rubra	97	Alive	23906	Medium
11	18.6	1.9	Bright Green	Quercus macrocarpa	107	Alive	24492	High
12	18.6	1.56	Olive Green	Quercus velutina	83	Alive	29730	Medium

- What are the columns?

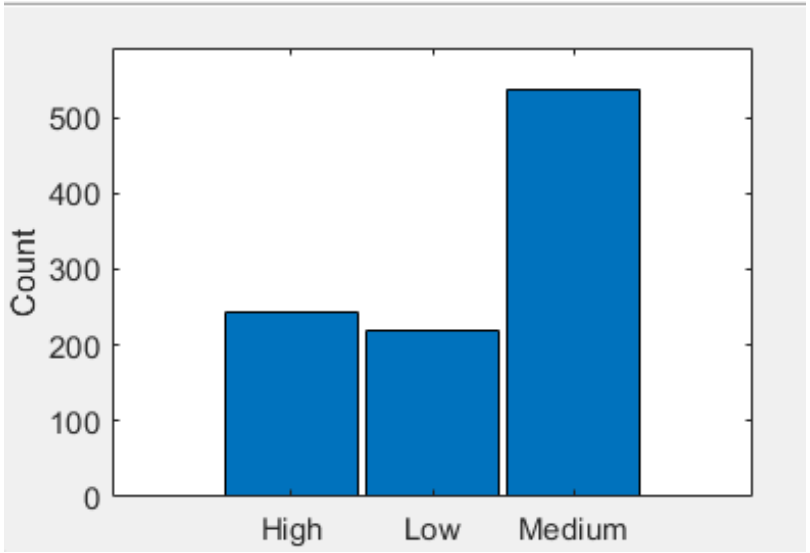
Data Types-Example Models (Oak Tree)

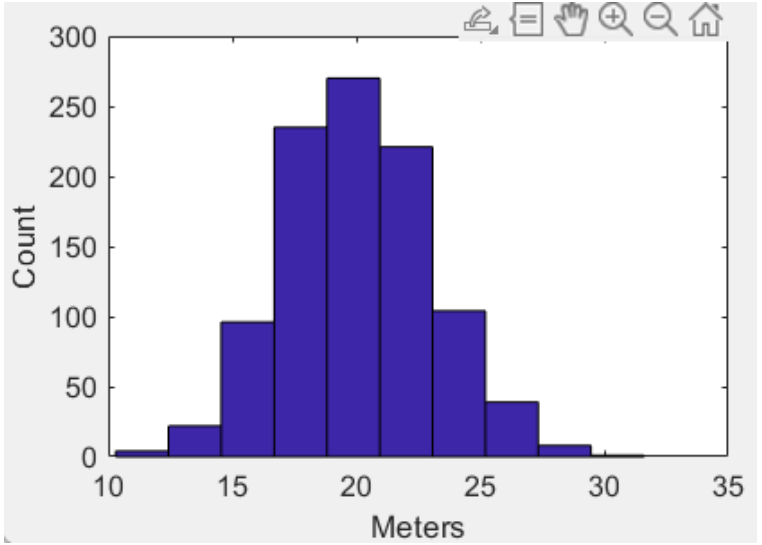
- **Nominal (Categorical, Unordered)**
- *Example:* Species (Quercus rubra, Quercus alba, etc.)
- *Modeling:* Classification (Decision Tree, Random Forest)
- *Example Heuristic:* Predict species from height, width, and leaf color.
- **Ordinal (Categorical, Ordered)**
- *Example:* Lumber Value (Low, Medium, High)
- *Modeling:* Ordinal Logistic Regression
- *Example Heuristic:* Higher lumber value if height × width > 35, medium if > 25.
- **Boolean (Binary)**
- *Example:* Alive/Dead
- *Modeling:* Logistic Regression or Gradient Boosted Classifier
- *Example Heuristic:* Probability of being alive increases with age < 100 years and wid

Data Types-Example Models (Oak Tree)

- **Discrete Numeric**
- *Example:* Number of Leaves
- *Modeling:* Poisson or Negative Binomial Regression
- *Example Heuristic:* Leaf count ≈ 500 × age (if alive), reduced by 90% if dead.
- **Continuous Numeric**
- *Example:* Height (m)
- *Modeling:* Linear Regression
- *Example Heuristic:* Height increases ~0.3 m per year until plateau at ~25 m.
- **Datetime / Time Series**
- *Example:* Growth measurements over years
- *Modeling:* ARIMA, SARIMA, or Prophet
- *Example Heuristic:* Annual growth shows seasonal spikes in spring and summer.

MATLAB!





- Early visualization and summary statistics