CS109A: Introduction to Data Science
Lecture 02: Data and Visualization

Harvard University

Fall 2024

- **Course:** CS109A: Introduction to Data Science
- **Lecture:** Lecture 02: Data and Visualization
- **Instructor:** Kevin Rader
- **Objective:** Understand data types, collection methods, descriptive statistics, and visualization techniques for exploratory data analysis

Key Summary

This lecture covers the fundamental building blocks of data science: understanding what data are, where they come from, how to structure them for analysis, and how to explore them effectively. We cover the distinction between populations and samples, measures of center (mean, median, mode) and spread (variance, standard deviation), and the critical importance of data visualization. The famous Anscombe's Quartet demonstrates why you should **always visualize your data** rather than relying solely on summary statistics. We also explore various visualization techniques including histograms, bar plots, scatter plots, box plots, and violin plots.

Contents

1 Key Terminology

Before diving into data analysis, you need to understand these fundamental terms:

Table 1: *Essential Data Science Terminology*

Term	Simple Explanation	Notes
Data	A collection of facts, values, or information obtained through observation	Plural form; singular is "datum"
Tabular Data	Data organized in rows and columns like a spreadsheet	Also called "Tidy Data"
Observation	A single unit of analysis (one row)	Example: one movie, one student
Variable	A characteristic being measured (one column)	Also called "feature" or "attribute"
Population	The entire group you want to study	Example: ALL students in this class
Sample	A subset taken from the population	Example: Students who attended today
EDA	Exploratory Data Analysis	Using visuals and statistics to find patterns
Measures of Center		
Mean ($\bar{x}$)	Sum of all values divided by count	Sensitive to outliers
Median	The middle value when data is sorted	Robust to outliers
Mode	Most frequently occurring value	Used for categorical data
Measures of Spread		
Variance (s^2)	Average squared distance from the mean	Units are squared
Std. Dev. (s)	Square root of variance	Same units as original data

2 What is Data?

Data science starts with **data**. But what exactly is data?

2.1 Definition of Data

Definition:

Data vs. Datum

- **Datum** (singular): A single piece of information or measurement
- **Data** (plural): A collection of information pieces obtained through observation or measurement

Data can be:

- **Numeric**: Numbers (integers, decimals)
- **Categorical**: Groups or categories ("male/female", "red/blue/green")
- **Boolean**: True/False, Yes/No, 1/0
- **Strings**: Text ("Hello World")

In the modern world, **everything is data**. Facebook collects your social interactions, Google tracks your searches, your phone records your location, and even your grocery store tracks your purchases.

2.2 Where Does Data Come From?

Data can come from three main sources:

2.2.1 1. Internal Sources (Primary Data)

Data you or your organization collect directly:

- Scientific experiments
- Clinical trials
- Surveys you design and distribute
- Company sales records

2.2.2 2. Existing External Sources

Data that someone else has already collected and made available:

- Government open data portals
- Kaggle datasets
- Sports statistics websites
- Academic research databases

2.2.3 3. External Sources Requiring Collection

Data that exists online but requires effort to extract:

- **APIs**: Official interfaces provided by companies
- **RSS Feeds**: Streams from blogs and news sites
- **Web Scraping**: Extracting data from HTML pages

2.3 Methods of Online Data Collection

Table 2: *Three Methods of Online Data Collection*

Method	Description	Pros/Cons
API	Official interface provided by companies (Google Maps, Twitter, Spotify)	Pro: Legal, stable, accurate data. Con: Often rate-limited or paid.
RSS	Streams of updated content from blogs/news sites	Pro: Free, real-time updates. Con: Limited to what publishers provide.
Web Scraping	Extracting data directly from HTML code	Pro: Flexible, free. Con: Legal/ethical concerns, fragile.

Warning

Ethical and Legal Concerns with Web Scraping

Web scraping is powerful but comes with serious responsibilities:

- **Terms of Service**: Many websites explicitly prohibit scraping
- **Privacy**: Never collect or publish private/personal information
- **Server Load**: Excessive scraping can overwhelm servers (similar to a DoS attack)
- **Potential Harm**: Ask yourself: "Could publishing this data harm someone?"

Rule of thumb: Just because data is publicly available doesn't mean you can use it however you want. Always ask "Should I be doing this?" before scraping.

3 Data Types and Structures

3.1 Atomic Data Types

The most basic building blocks of data:

- **Numeric**:
  - **Integers**: Whole numbers (e.g., 42, -7, 0)
  - **Floats**: Decimal numbers (e.g., 3.14, -0.001)
- **Boolean**: True/False values (1/0, Yes/No)
- **Strings**: Text sequences (e.g., "Hello", "CS109A")

3.2 Compound Data Types

More complex structures built from atomic types:

- **Lists**: Ordered sequences of values

```
my_list = [1, 2, 3, 4, 5]
```

- **Dictionaries**: Key-value pairs

```
student = {
    "first": "Kevin",
    "last": "Rader",
    "classes": ["CS109A", "STAT104"]
}
```

3.3 Tabular Data: The Gold Standard

Key Information

Why Tabular Data Matters

Most data analysis tools (pandas, sklearn, etc.) expect your data in **tabular format**—like an Excel spreadsheet with rows and columns.

- **Rows = Observations**: Each row represents one unit of analysis (one movie, one student, one transaction)
- **Columns = Variables**: Each column represents one characteristic being measured (rating, age, price)

```
import pandas as pd

# Read a CSV file into a DataFrame
imdb = pd.read_csv('imdb_top_1000.csv')

# View the first 5 rows
imdb.head()
```

Listing 1: Loading tabular data with pandas

The output shows a table where:

- Each **row** is a different movie (The Shawshank Redemption, The Godfather, etc.)
- Each **column** is a variable (Series_Title, Released_Year, IMDB_Rating, etc.)

3.4 Variable Types: A Critical Distinction

Important:

Why Variable Types Matter The type of variable determines:

- Which **summary statistics** you can calculate
- Which **visualizations** are appropriate
- Which **models** you can use

You can calculate the mean of "height" but not the mean of "favorite color"!

Variables are divided into two main categories:

3.4.1 Quantitative (Numeric) Variables

Values are numbers where arithmetic operations make sense.

- **Discrete**: Values are countable, typically integers
  - Examples: Number of siblings, dice rolls, page views
- **Continuous**: Values can be any number within a range
  - Examples: Height, weight, temperature, time

3.4.2 Categorical Variables

Values fall into distinct groups or categories.

- **Nominal**: No natural ordering
  - Examples: Blood type (A, B, O, AB), pet preference (dog, cat, rat)
- **Ordinal**: Natural ordering exists
  - Examples: Letter grades (A, B, C, D, F), satisfaction (high, medium, low)

3.5 Messy Data and Tidy Data

Real-world data is rarely clean. Common problems include:

- **Missing values**: Empty cells
- **Wrong values**: Data entry errors (age = 200)
- **Format mismatches**: Different date formats, inconsistent naming
- **Messy structure**: Data not in tabular format

Example:

Converting Messy Data to Tidy Data **BEFORE (Messy):** Weekend produce deliveries

	Friday	**Saturday**	**Sunday**
Morning	15	158	10
Afternoon	2	90	20
Evening	55	12	45

Problems:

- One row contains 3 observations (Friday morning, Saturday morning, Sunday morning)
- "Friday" is a value, not a variable name
- Hard to calculate "average deliveries" or "total by day"

AFTER (Tidy):

ID	**Time**	**Day**	**Deliveries**
1	Morning	Friday	15
2	Morning	Saturday	158
3	Morning	Sunday	10
4	Afternoon	Friday	2
5	Afternoon	Saturday	90
...	...	...	...
9	Evening	Sunday	45

Why this is better:

- **1 observation = 1 row**: Each row is a unique time-day combination
- **1 variable = 1 column**: Time, Day, and Deliveries are separate columns
- Easy to filter, group, and analyze with pandas

4 Population vs. Sample

Understanding the distinction between population and sample is fundamental to all of statistics.

Definition:

Population and Sample

- **Population**: The *entire* set of objects/individuals you want to study
  - Example: ALL students enrolled in CS109A
- **Sample**: A *subset* of the population that you actually observe
  - Example: Students who attended class today

We analyze the **sample** to make inferences about the **population**.

4.1 Sampling Bias

The sample must be **representative** of the population. When it isn't, we have **sampling bias**.

Table 3: *Types of Sampling Bias*

Type	Description
Selection Bias	Some individuals are more likely to be selected than others
Non-response Bias	People who don't respond may be systematically different from those who do
Volunteer Bias	People who volunteer may be more enthusiastic or have stronger opinions

Example:

Sampling Bias in Practice **Example 1: Class Attendance**

- **Goal**: Survey satisfaction of ALL CS109A students
- **Sample**: Only students who came to class today
- **Problem**: Students who skip class might be less satisfied—their opinions are missing!
- **Result**: Overestimated satisfaction scores

Example 2: Early Adopters

- **Goal**: Test if a new app feature works for all users
- **Sample**: "Early adopters" who signed up for beta testing
- **Problem**: Early adopters are tech-savvy and excited about new features
- **Result**: Feature seems great in testing, but flops when released to everyone

Lesson: Always ask "Who is missing from my sample?" and "How might they be different?"

5 Descriptive Statistics: Measures of Center

Descriptive statistics summarize the characteristics of your data. We start with measures of **center**—where the "typical" value is located.

5.1 Mean (Average)

Definition:

Sample Mean The **mean** is the sum of all values divided by the count:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Intuition: The mean is the "balancing point" of the distribution—if you placed the data on a seesaw, the mean is where it would balance.

Warning

The Mean is Sensitive to Outliers

Consider these two datasets:

- Dataset A: [1, 2, 3, 4, 5] → Mean = 3
- Dataset B: [1, 2, 3, 4, **100**] → Mean = 22

One extreme value (100) pulled the mean from 3 to 22! This is why the mean can be misleading when outliers are present.

5.2 Median

Definition:

Sample Median The **median** is the middle value when data is sorted:

- If n is odd: The middle value
- If n is even: Average of the two middle values

Advantage: The median is **robust** to outliers—extreme values don't affect it much.

Example:

Mean vs. Median Student ages: [17, 19, 21, 22, 23, 24, 26, **38**]

Median: Average of 22 and 23 = **22.5**

Mean: $(17 + 19 + 21 + 22 + 23 + 24 + 26 + 38)/8 = \mathbf{23.75}$

The 38-year-old graduate student pulls the mean up, but the median barely moves!

5.3 Mean vs. Median: Skewness

The relationship between mean and median tells you about the **shape** of the distribution:

Table 4: *Mean, Median, and Distribution Shape*

Distribution Shape	Relationship	Example
Symmetric	Mean $\approx$ Median	Normal (bell curve)
Right-skewed	Mean $>$ Median	Income distribution
Left-skewed	Mean $<$ Median	Age at retirement

Key Information

Why Financial Data Uses Medians

Income and housing prices are typically **right-skewed**—most people earn moderate amounts, but a few billionaires earn enormously more.

If we reported mean income, those billionaires would make everyone seem richer than they are. That's why economists report **median household income**—it better represents the "typical" household.

5.4 Mode

Definition:

Mode The **mode** is the most frequently occurring value in a dataset.

Use case: Primarily for **categorical data** where mean and median don't make sense.

Example:

Mode for Categorical Data Favorite pets: ["Dog", "Cat", "Dog", "Rat", "Cat", "Dog"]

Mode: "Dog" (appears 3 times)

You can't calculate the "mean" of pet preferences, but you can find the most popular choice!

5.5 Computational Efficiency

Key Information

Mean is Faster to Compute Than Median

- **Mean**: $O(n)$ — Just scan through once, keeping track of sum and count
- **Median**: $O(n \log n)$ — Requires sorting first!

When you have billions of data points (like Facebook or Google), this difference matters. This is one reason why means are more commonly reported even when medians might be more appropriate.

6 Descriptive Statistics: Measures of Spread

Knowing the center isn't enough—we also need to know how **spread out** the data is.

6.1 Range

Definition:

Range

$$\text{Range} = \text{Maximum} - \text{Minimum}$$

Problem: Only uses two values (max and min), ignoring everything in between.

6.2 Variance

Definition:

Sample Variance The **variance** measures the average squared distance from the mean:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Steps:

1. Calculate the mean ($\bar{x}$)
2. For each value, find the distance from the mean ($x_i - \bar{x}$)
3. Square each distance (removes negatives, emphasizes larger deviations)
4. Take the average of squared distances (dividing by $n-1$)

Example:

Why $n-1$ Instead of n? **Question**: Why do we divide by $n-1$ instead of n?
Simple Intuition: What's the minimum number of observations needed to measure "spread"? With just **one observation** (e.g., [5]), you can't say anything about how spread out the data is! The formula with $n-1$ agrees: $\frac{1}{1-1} = \frac{1}{0}$ is undefined.
You need at least 2 observations to talk about spread.
Technical Answer: Dividing by $n-1$ (called "degrees of freedom") corrects for a bias that occurs because we estimate the mean from the same data. This gives us an "unbiased estimator" of the true population variance.

6.3 Standard Deviation

Definition:

Sample Standard Deviation

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Key advantage: The standard deviation has the **same units as the original data**, making it interpretable.

If heights are measured in cm, variance is in cm^2 (meaningless), but standard deviation is back in cm!

Key Information

Interpreting Standard Deviation

The standard deviation is roughly the "average distance" from the mean.

If the mean height is 170 cm and the standard deviation is 10 cm, most people are within about 10 cm of 170 cm (between 160–180 cm).

7 Why Visualization Matters: Anscombe's Quartet

Important:

Never Trust Summary Statistics Alone! **Anscombe's Quartet** is a famous example showing why you must **always visualize your data**.

Four different datasets have:

- Same mean of X (9.0)
- Same mean of Y (7.50)
- Same variance of X (11.0)
- Same variance of Y (4.12)
- Same correlation (0.816)

Based on these statistics, the datasets seem identical. But when you plot them...

The Four Datasets (with identical summary statistics):

Dataset I		Dataset II		Dataset III		Dataset IV	
X	Y	X	Y	X	Y	X	Y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

What the scatter plots reveal:

- **Dataset I**: Linear relationship with random scatter (what you'd expect)
- **Dataset II**: Perfect **parabola**—clearly nonlinear!
- **Dataset III**: Perfect line except for one **Y-outlier**
- **Dataset IV**: All X values are 8, except one **X-outlier** at 19

Key Summary

The Lesson of Anscombe's Quartet

Always visualize your data before modeling. Summary statistics can hide crucial patterns:

- Nonlinear relationships
- Outliers
- Clusters or subgroups
- Data quality issues

Don't just calculate R^2 and call it a day!

8 Basic Visualization Types

Different visualizations serve different purposes. Choose based on what you want to learn.

8.1 Visualizations by Purpose

Table 5: *Choosing the Right Visualization*

Chart Type	Purpose	Variables	What to Look For
Histogram	Distribution of one numeric variable	1 numeric	Shape, center, spread, outliers
Bar Plot	Frequency of categories	1 categorical	Which categories are most/least common
Scatter Plot	Relationship between two variables	2 numeric	Direction, strength, form, outliers
Box Plot	Compare distributions across groups	1 numeric + 1 categorical	Median, IQR, outliers by group
Violin Plot	Compare distribution shapes across groups	1 numeric + 1 categorical	Full distribution shape (bimodality, etc.)
KDE Plot	Smooth distribution curve	1 numeric	Avoids bin-width problems of histograms

8.2 Histogram vs. Bar Plot

Both use bars, but they're fundamentally different:

Table 6: *Histogram vs. Bar Plot*

Feature	Histogram	Bar Plot
Variable type	Numeric (continuous)	Categorical
Bars	Touch each other	Separated
X-axis	Has natural order	Order is arbitrary
Purpose	Show distribution shape	Compare category frequencies

8.3 Histogram Bin Width Matters

Warning

The appearance of a histogram depends heavily on **bin width**:

- **Too wide**: Loses detail, everything looks uniform
- **Too narrow**: Too much noise, hard to see patterns
- **Just right**: Shows the true shape of the distribution

Always try multiple bin widths to make sure you're not missing important patterns!

Alternative: Use a **Kernel Density Estimate (KDE)** which creates a smooth curve and avoids the bin-width problem.

8.4 Scatter Plots: Reading Relationships

When looking at a scatter plot, ask four questions:

1. **Direction**: Is the relationship positive (both increase together) or negative (one increases as the other decreases)?
2. **Strength**: How tightly do points cluster around the trend? (Strong = tight, Weak = scattered)
3. **Form**: Is the relationship linear or curved?
4. **Outliers**: Are there any points that don't fit the pattern?

8.5 Box Plots: Comparing Distributions

A box plot shows five key statistics:

- **Median**: Line in the middle of the box
- **Q1 and Q3**: Bottom and top of the box (middle 50% of data)
- **Whiskers**: Extend to the smallest/largest non-outlier values
- **Outliers**: Individual points beyond the whiskers

Use case: Comparing how a numeric variable differs across categories (e.g., income by education level).

8.6 Violin Plots: More Detail Than Box Plots

A violin plot is a box plot combined with a KDE:

- Shows the full distribution shape on both sides
- Reveals bimodality or unusual shapes that box plots hide
- More informative but takes more space

8.7 Why Pie Charts Are Usually Bad

Warning

The Problem with Pie Charts
Humans are bad at comparing angles. When pie slices are similar sizes, it's nearly impossible to tell which is bigger.
Better alternative: Use a bar plot. Humans easily compare bar heights.
Exception: Pie charts work when you want to emphasize "parts of a whole" and there are only 2–3 categories with very different sizes.

9 Visualizing Multiple Variables

What if you have more than two variables? Three-dimensional scatter plots are tempting but usually fail on 2D screens.

9.1 The Problem with 3D Scatter Plots

A "scatter cloud" is nearly impossible to interpret:

- Depth perception is lost on a flat screen
- Occlusion: points hide behind other points
- You can't rotate the view interactively in a static report

9.2 Better Approaches: Aesthetic Mappings

Instead of adding a third spatial dimension, map additional variables to visual properties:

- **Color**: Great for categorical variables (up to 5–8 groups)
- **Size**: Good for numeric variables
- **Shape**: Good for categorical variables (2–3 groups max)
- **Animation**: Good for time (each frame = one time point)

Example:

Gapminder: 5 Variables in One Plot The famous "Wealth and Health of Nations" visualization shows:

1. **X-axis**: Income per person
2. **Y-axis**: Life expectancy
3. **Circle size**: Population
4. **Circle color**: Continent
5. **Animation**: Year (time)

Key insights visible:

- Positive (but nonlinear) relationship between income and life expectancy
- China's dramatic rise from poverty to prosperity
- The US has lower life expectancy than peers with similar income

Design choices:

- Population → Size (numeric → size works well)
- Continent → Color (categorical → distinct colors)

10 Historical Examples of Data Visualization

Data visualization isn't new—some of history's most important discoveries came from visualizing data.

10.1 John Snow's Cholera Map (1854)

- **Problem**: Cholera outbreak in London—source unknown
- **Visualization**: Dot map showing location of each cholera death
- **Discovery**: Deaths clustered around the Broad Street water pump
- **Action**: Removed the pump handle; outbreak ended
- **Legacy**: Proved that cholera spread through contaminated water, not "bad air"

10.2 Florence Nightingale's Rose Chart (1858)

- **Problem**: High death rates in military hospitals during the Crimean War
- **Visualization**: Rose chart showing deaths by cause over time
- **Discovery**: Blue (preventable disease) vastly exceeded red (combat wounds)
- **Action**: Hospital sanitation reforms
- **Legacy**: Pioneered the use of statistics in healthcare policy

10.3 Charles Minard's Napoleon Map (1869)

- **Subject**: Napoleon's disastrous invasion of Russia (1812)
- **Variables shown**: Army size (line width), geography (map), direction (color), time, and temperature
- **Story**: 422,000 troops entered; only 10,000 returned
- **Legacy**: Called "the best statistical graphic ever drawn"

11 Principles of Effective Visualization

11.1 1. Graphical Integrity

Don't lie with your visuals.

- Y-axis should usually start at zero for bar charts
- Geographic maps should account for population, not just area
- Don't use 3D effects that distort perception

11.2 2. Keep It Simple

Avoid "chart junk"—unnecessary decoration that doesn't help understanding.

- No 3D effects on 2D data
- No excessive gridlines or backgrounds
- Maximize the "data-ink ratio" (information per pixel)

11.3 3. Use the Right Display

Visual channels have different effectiveness for encoding data:

1. Position (most accurate)
2. Length
3. Angle (pie charts are here—not great!)
4. Area
5. Color intensity
6. Shape (least accurate for quantitative data)

11.4 4. Use Color Strategically

- **Qualitative**: Distinct colors for categories (limit to 5–8)
- **Sequential**: Light-to-dark for ordered values
- **Diverging**: Two colors from a neutral center (e.g., blue–white–red)
- **Avoid**: Rainbow color maps (no natural order)
- **Consider**: Color blindness (avoid red-green combinations)

11.5 5. Know Your Audience

- **Exploratory**: For yourself—neutral, finding patterns
- **Explanatory**: For others—making a specific point

12 Key Takeaways

Key Summary

Summary of Lecture 02

What is Data?

- Data = collected information; tabular format is ideal
- Variables can be numeric (discrete/continuous) or categorical (nominal/ordinal)
- Always consider: Where did this data come from? What might be missing?

Population vs. Sample

- Population = everyone you care about
- Sample = the subset you actually observe
- Watch out for sampling bias (selection, non-response, volunteer)

Measures of Center

- Mean: Balancing point (sensitive to outliers)
- Median: Middle value (robust to outliers)
- Mode: Most frequent (for categorical data)
- If Mean > Median: Right-skewed distribution

Measures of Spread

- Variance: Average squared deviation (units are squared)
- Standard Deviation: Square root of variance (same units as data)

Visualization

- Anscombe's Quartet: ALWAYS visualize before modeling!
- Choose the right chart for your purpose
- Histograms/KDE for distributions, scatter plots for relationships, box/violin plots for comparisons
- Multiple variables: Use color, size, shape, animation
- Follow principles: integrity, simplicity, appropriate display, strategic color