From Questions to Data

Measurement and Exploratory Data Analysis

Sergio I. García-Ríos

Fall 2026

PA 311N: Data Analysis for Public Policy
LBJ School of Public Affairs

What are we trying to do?

Public-policy decisions are often arguments about evidence.

Questions we might ask

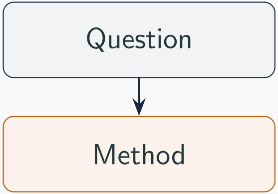
- Is housing becoming less affordable?
- Are crime rates increasing?
- Does a program improve student outcomes?
- Do voters support a policy?

The harder question

What evidence would we need to answer these well?

Data analysis starts with a question

- **What** do we want to know?
- **Who or what** are we studying?
- What would we need to **observe**?
- What evidence would **change our mind**?



Methods should follow the question, not the other way around.

From the world to a dataset

Real world → Measurement → Dataset → Explore → Conclusion

A dataset is a simplified representation of the world.

Every dataset reflects choices about what gets included, what gets measured, how things are defined, and what gets left out.

The basic building blocks

- Observations / cases: people, cities, schools, countries, years
- Variables: characteristics measured for each case
- Values: what we observe for each variable

City	Income	Unemp.	Region
Austin	92,000	3.4	South
El Paso	57,000	4.6	South
Lubbock	63,000	3.8	South

Rows are observations. Columns are variables. Each cell contains a value.

Population and sample

Population

Everyone or everything we want to understand.

Sample

The cases we actually observe.

Example: We want to understand **Texas voters**, but interview **1,200 Texas voters**.

Sampling is something we already do

Imagine tasting one spoonful of soup.

- Spoonful = sample
- Whole pot = population
- Deciding whether the pot needs salt = inference

For an inference to work, the sample needs to tell us something useful about the population.

Measurement: what exactly are we measuring?

Is Austin becoming less affordable?

What does affordable mean?

- Median rent
- Home prices
- Rent as a share of income
- Cost-burdened households
- Homelessness
- Income after housing costs

Key point

Different measures can produce different answers to the same substantive question.

Variables are representations

Education	Income	Ideology
Years of schooling	Individual income	Liberal–conservative
Highest degree	Household income	Party identification
College graduate: yes/no	Income categories	Policy preferences

There is rarely one inevitable way to measure a concept.

Two broad types of variables

Categorical

Values represent groups or labels.

- Party
- Region
- Employment status

Quantitative

Values represent quantities.

- Age
- Income
- Unemployment rate

Variable type tells us how to summarize and visualize the data.

Exploratory Data Analysis

Before we model, we learn what is in the data.

Start with one variable

For any variable, ask three questions:

1. What values occur?
2. How often do they occur?
3. What does the distribution look like?

The first goal of EDA is description, not explanation.

Categorical variables: count and proportion

Imagine a hypothetical survey of 100 residents asks whether they support expanding public transit.

Response	Count	Proportion
Support	52	0.52
Oppose	31	0.31
Unsure	17	0.17
Total	100	1.00

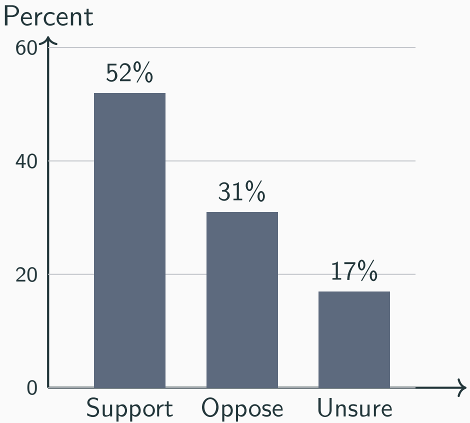
Two useful summaries

Count: how many observations are in each category.

Proportion: the share of observations in each category.

52 out of 100 $= 0.52 =$ 52%.

Categorical variables: visualize with a bar plot



Why a bar plot?

The height of each bar represents the frequency or proportion in a category.

Your turn

What is the main pattern in one sentence?

Quantitative variables: look at the distribution

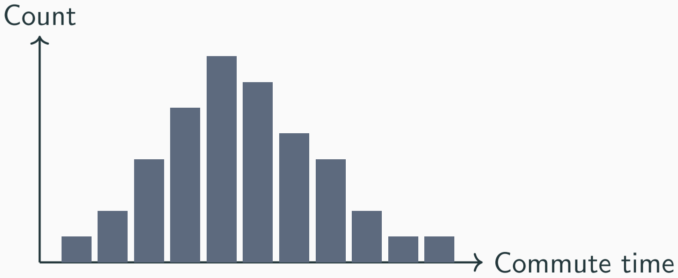
A quantitative variable has many possible values, so counts alone are usually not enough.

We want to know:

- Where are most observations?
- How spread out are they?
- Is the distribution symmetric or skewed?
- Are there unusual observations?

A histogram is often the first place to look.

A histogram groups quantitative values



Read the shape

Most observations are clustered in the middle, with fewer long commutes.

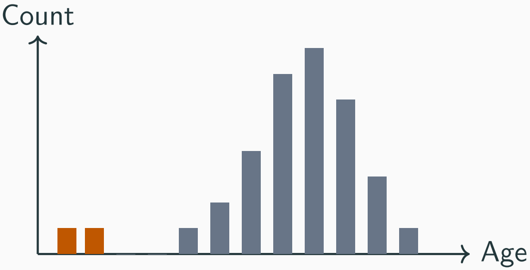
The exact bars depend on how the values are grouped into bins.

Something looks strange

Imagine a class survey asks:

"How old were you when you had your first kiss?"

Most answers fall between 12 and 20, but several students report ages 0–3.



What might be going on?

The question may be ambiguous: a romantic kiss, or any kiss?

Measuring the center: the mean

The mean is the arithmetic average.

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Example commute times, in minutes:

10, 12, 15, 18, 20, 22, 25, 27, 30, 120

Mean commute time

29.9 minutes

Measuring the center: the median

The median is the middle value after ordering the observations.

$$10, 12, 15, 18, 20, 22, 25, 27, 30, 120$$

With an even number of observations, take the average of the two middle values.

$$\text{Median} = \frac{20 + 22}{2} = 21$$

Why is the mean 29.9 but the median only 21?

Mean or median? Look at the distribution

Mean	Median
Uses every value.	Uses the ordering of values.
Sensitive to extreme observations.	Resistant to extreme observations.
Often useful for roughly symmetric distributions.	Often useful for skewed distributions.

The “best” summary depends on what the distribution looks like.

Center is not enough: we also need spread

Two neighborhoods can have the same average commute but very different experiences.

Neighborhood A	Neighborhood B
24, 25, 25, 26, 25	5, 15, 25, 35, 45
Mean $= 25$	Mean $= 25$

Same center. Very different spread.

Two simple measures of spread

Range

Maximum − minimum

For our commute example:

$$120 - 10 = 110$$

Interquartile range (IQR)

Spread of the middle 50% of observations.

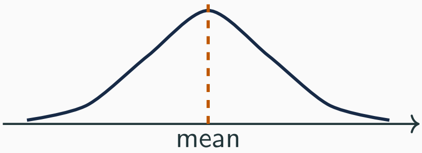
$$IQR = Q_3 - Q_1$$

Like the median, the IQR is less sensitive to extreme values.

Standard deviation: a preview

The standard deviation summarizes how far observations typically fall from the mean.

- Small SD: values cluster near the mean.
- Large SD: values are more dispersed.
- It is expressed in the same units as the variable.



We will develop the standard deviation more carefully when we study distributions and probability.

Match the variable to the summary

Variable	Useful numerical summary	Useful visualization
Categorical	Count, proportion / percentage	Bar plot
Quantitative	Mean or median + measure of spread	Histogram, boxplot

A good habit

Do not report a mean without first looking at the distribution that produced it.

Before you trust a summary, inspect the data

Look for

- Missing values
- Impossible values
- Unexpected categories
- Extreme observations

Ask why

A strange value could be:

- real,
- a data-entry error,
- a coding problem,
- a measurement problem.

Investigate first. Do not automatically delete unusual observations.

What can the evidence support?

Description	Association	Causation
What is happening?	**Are two things related?**	**Does X change Y?**
What percentage of Texans lack health insurance?	Is income related to insurance coverage?	Does expanding Medicaid increase coverage?

Today we focused on description. Association and causation require additional tools.

The workflow we start this week

Question → Data → Explore → Summarize → Interpret → Communicate

Wednesday in R

We will identify variable types, inspect data, calculate counts/proportions and simple numerical summaries, and create a visualization in a reproducible document.

If I give you a variable, how should you summarize and visualize it?