ECON 0150 | Economic Data Analysis | Course Outline

This document describes the structure and content of the course. It serves as a reference for students navigating the material, instructors planning sessions, and anyone seeking to understand the course design.

Part 0: Framework

The conceptual toolkit for the entire course. Everything that follows builds on these foundations.

What is data?

Data is the realization of a random variable we cannot directly observe. We're interested in the random variable itself—the underlying process generating the data—not the particular data points we happen to have. The data is just what we have to work with.

This perspective matters: our goal is always to learn about the process, not to describe our specific sample. This idea returns in Part 3 when we formalize inference.

How do we organize data?

We write **x_it** where **i** indexes entities and **t** indexes time.

This notation distinguishes two types of variables:

- **Substantive variables** — the phenomena being measured (income, price, sentiment)
- **Index variables** — the structure organizing observations (i for entities, t for time)

Different **data structures** emerge from which indices are active:

Structure	What varies	What's fixed	Example
Cross-section	entities (i)	time (t)	Income across households in 2024
Time series	time (t)	entity (i)	US GDP from 1950–2024
Panel	both i and t	—	Income across households, 2020–2024
Geographic	location	—	Unemployment by county

The Building Blocks

Throughout the course, you accumulate expertise in four areas. Each unit adds to at least one.

Variables — *What kind of values?*

- Categorical: binary, nominal, ordinal
- Numerical: discrete, continuous

Structures — *How is it organized?*

- Cross-section, time series, panel, geographic

Operations — *What do I do to it?*

- Filter, count, bin, group, reshape, merge, transform (log, difference, adjust for inflation)

Visualizations — *How do I show it?*

- Bar, pie, histogram, boxplot, line, scatter, multi-line, heatmap, map

How visualizations connect to structure

The choice of visualization depends on how many substantive variables you're analyzing and which index dimensions are active.

What you're showing	Indexes active	Substantive vars	Example visualization
Distribution of x	i implicit	1	Histogram, boxplot
x over time	t	1	Line plot
x across groups	i as category	1	Grouped boxplot
x vs y	i implicit	2	Scatterplot
x across entities and time	i and t	1	Multi-line plot (panel)
x vs y across entities	i and t	2	Small multiples, heatmap

Key insight: A time series plot has two axes (value and time), but it's still univariate analysis. Time is structural scaffolding, not a substantive variable—you're describing one phenomenon's trajectory, not analyzing the relationship between GDP and time the way you'd analyze GDP and unemployment.

Similarly, panel data doesn't add more substantive variables—it makes both index dimensions analytically active. The i subscript stops being a row label and becomes something you can visualize against (country, firm, person).

How we work with data

Every visualization follows three steps:

1. **SELECT** — What subset of data are we looking at?
2. **TRANSFORM** — How do we summarize or reshape it?
3. **ENCODE** — How do we map values to visual elements?

This SELECT-TRANSFORM-ENCODE pattern repeats throughout the course. Each unit introduces new operations and visualizations, but the workflow remains the same.

Course progression

The course builds complexity along two axes:

Axis	Progression
Data structure	Cross-section → Time series → Panel
Analysis type	Single variable → Relationships → Statistical models

Parts 1–2 focus on exploratory analysis: describing variables and visualizing relationships.

Parts 3–5 focus on statistical modeling: making inferences about populations from samples.

How each class works

Each unit follows a consistent rhythm:

When	What	Purpose
Before class	Concept video (~10 min)	Learn the core ideas at your own pace
Start of class	Quiz	Confirm you watched; replaces attendance
During class	Exercise	Guided practice with support
After class	Homework	Independent practice

Exercises are homework prep. The exercise and homework use the same skills, but:

- **Exercise:** Done in class with support (instructor, UTAs, peers). It's okay to struggle and ask questions.
- **Homework:** Done independently. Similar structure to the exercise. Demonstrates you've mastered the skill.

You should leave each class thinking: "I just practiced exactly what I need to do for homework."

Part 1: Exploring Variables

One substantive variable at a time. Cross-section first, then time series. Variable types are introduced in Part 0; Part 1 applies the framework to visualize data.

Unit	Structure	Focus	Visualizations	Description
1.1	Cross-section	Categorical	Bar chart, pie	Describe frequencies and proportions of a categorical variable.
1.2	Cross-section	Numerical	Histogram, boxplot	Describe distribution, center, and spread of a numerical variable.
1.3	Time series	Numerical	Line plot	Describe a variable over time; introduce trend and seasonality.
1.4	Time series	Transformations	Line plot	Apply log, inflation adjustment, and differencing; interpret transformed series.

After Part 1, you have:

- Variables: binary, nominal, ordinal, discrete, continuous
- Structures: cross-section, time series
- Operations: count, bin, quartiles, filter, sort, log, inflation adjust, difference
- Visualizations: bar, pie, histogram, boxplot, stripplot, line

Exam 1

Part 2: Exploring Relationships

Adding complexity: a second substantive variable, or activating the second index dimension (panel).

Why panel belongs here: Panel data makes both i and t analytically active. Like adding a second substantive variable, this involves asking "how does this vary with that?"—sometimes across variables, sometimes across the i × t structure.

Long vs wide format: This distinction matters when both i and t are active. Format determines which visualizations are natural.

Format	What's a row?	Natural for...
Long	One (i, t) observation	Multi-line plots, facets by entity, scatterplots
Wide	One entity i	Heatmaps, side-by-side comparisons across time

Unit	Structure	Focus	Visualizations	Description
2.0	—	Framework	Anscombe's quartet	Motivate why we visualize relationships; correlation isn't everything.
2.1	Cross-section	Num × Num	Scatter	Explore relationships between two numerical variables; correlation.
2.2	Cross-section	Num × Cat	Grouped boxplot	Compare a numerical variable across categories.
2.3	Panel	Structure	Multi-line plot	Introduce panel as cross-section + time series; both i and t active.
2.4	Panel	Format & Viz	Heatmap, small multiples	Long vs wide format; how layout determines visualization.
2.5	Panel	Relationships	Scatter by group	Examine relationships between variables within panel structure.
2.6	Geographic	Location as index	Maps	Visualize data with a spatial dimension.

After Part 2, you have:

- Variables: (no new types)
- Structures: + panel, geographic
- Operations: + correlation, group, aggregate, reshape, merge
- Visualizations: + scatter, grouped boxplot, multi-line, heatmap, small multiples, map

Exam 2

Part 3: Univariate GLM

From description to inference. We formalize the idea that data is a sample from a population and develop tools to quantify uncertainty.

Unit	Title	Description
3.1	Random Variables	Data as repeated draws from a population random variable; connecting to Part 0's "what is data?"
3.2	Sampling & CLT	The Central Limit Theorem: the sample mean's distribution, regardless of population shape.
3.3	Confidence Intervals	Calculate and interpret confidence intervals using the t-distribution.
3.4	Hypothesis Testing	Conduct one-sample t-tests; calculate p-values; interpret statistical significance.
3.5	Simplest GLM	Build the first generalized linear model: $y = \beta_0 + \varepsilon$; test the intercept coefficient.

After Part 3, you can:

- Distinguish sample statistics from population parameters
- Quantify uncertainty with confidence intervals
- Test hypotheses about population means
- Fit and interpret a simple GLM

Exam 3

Part 4: Bivariate GLM

Models with one predictor. How does y change with x?

Unit	Title	Description
4.1	Numerical Predictors	Regression with a continuous predictor; interpreting slope coefficients.
4.2	Model Residuals	Checking model assumptions using residuals; diagnosing problems.
4.3	Categorical Predictors	Regression with a categorical predictor; the two-sample t-test as a special case.
4.4	Time Series Models	Regression through time; detecting and addressing autocorrelation.

After Part 4, you can:

- Fit and interpret bivariate regression models
- Diagnose model problems using residuals

- Handle both numerical and categorical predictors
- Recognize when time series structure requires special treatment

Exam 4

Part 5: Multivariate GLM

Models with multiple predictors. Controlling for confounds, modeling group differences, and testing interactions.

Unit	Title	Description
5.1	Categorical Controls	Fixed effects: controlling for unobserved group differences with multiple intercepts.
5.2	Interactions	Testing whether the relationship between x and y varies by group.
5.3	Numerical Controls	Including multiple continuous predictors; interpreting "holding constant."

After Part 5, you can:

- Control for confounding variables
- Model heterogeneous effects across groups
- Interpret coefficients in multiple regression
- Choose appropriate model specifications

Exam 5

Summary: The Arc of the Course

Part	Focus	Key question
0	Framework	What tools do we need?
1	Variables	What does this variable look like?
2	Relationships	How do these variables relate?
3	Univariate GLM	What can we infer about the population?
4	Bivariate GLM	How does y change with x?
5	Multivariate GLM	How does y change with x, controlling for z?

The building blocks accumulate throughout. By the end, you have a toolkit of variable types, data structures, operations, and visualizations—plus the statistical framework to move from description to inference.