Lecture 1-2: Research vs EDA

DATA 510: Data Science Capstone

Lucas P. Cordova, Ph.D.

2026-05-11

Contrasts exploratory data analysis with formal research questions for capstone work: when to wander the data, when to commit to a testable claim, common research-question types, and how the two modes reinforce each other without p-hacking.

Table of contents

1 Learning Objectives

1.1 Today's objectives

1.2 What you will learn today

By the end of this session, you will be able to:

1. **Contrast** exploratory data analysis with a capstone-grade research question (purpose, rigor, and deliverables).
2. **Classify** a vague idea into descriptive, inferential, predictive, or causal research aims (and know when exploratory work is the honest next step).
3. **Sketch** an EDA plan that feeds a predictive question without treating every plot as a hypothesis test.
4. **Explain** why research questions and EDA form a loop, not a linear checklist, in real projects.

. . .

Course connection: This supports your **project proposal** (clear question, data, and planned analysis) and the **Data-Driven Scrum** habit of separating open-ended backlog items from testable backlog items.

2 Part 1: Two modes of thinking

2.1 Research vs EDA

EDA (exploratory data analysis)

The art of *discovering* what is in the data: distributions, joins gone wrong, outliers, drift, and surprises you did not know to ask about yet.

Research

Disciplined *inquiry*: you articulate a problem, connect it to theory or prior work, and commit to analyses that can support or refute a claim about the world (or about a model of the world).

Tukey framed EDA as hypothesis generation, not confirmation. The capstone needs both: enough EDA to trust the data, enough research structure to finish.

2.2 EDA in one sentence

EDA is an **iterative** process for uncovering patterns, anomalies, and structure before you fully trust downstream modeling or inference.

. . .

If your only plan is "run XGBoost and see what happens," you still need EDA to learn whether labels, leakage, and joins make that plan meaningful.

2.3 Quick Quiz: Modes of work

You open a new client extract. Nobody documented the schema. What is the most honest first move?

```
A. Write the final thesis paragraph so stakeholders feel progress.
B. Fit a large neural net and tune until validation loss looks good.
C. Profile tables, keys, missingness, and a few sanity plots before locking a research questi
D. Pick the strongest correlation and build the deck around it.
```

. . .

C. EDA is how you earn the right to ask sharp research questions. A and B skip the reality check; D is how you accidentally narrate noise.

3 Part 2: What EDA is for

3.1 EDA: the art of data discovery

3.2 Why we actually do EDA

EDA helps you:

- **Understand structure:** keys, grain, time alignment, units, and encoding surprises.
- **Surface new questions:** anomalies and subgroups suggest hypotheses you would not brainstorm in a vacuum.
- **Choose methods:** if outcomes are extremely skewed, labels are sparse, or sensors drift, your modeling and evaluation plan should change.

. . .

EDA is not "fishing until p is small." It is **sense-making** and **risk reduction** for everything that follows.

3.3 Quick Quiz: EDA output

Which outcome is *most* aligned with healthy EDA for a capstone?

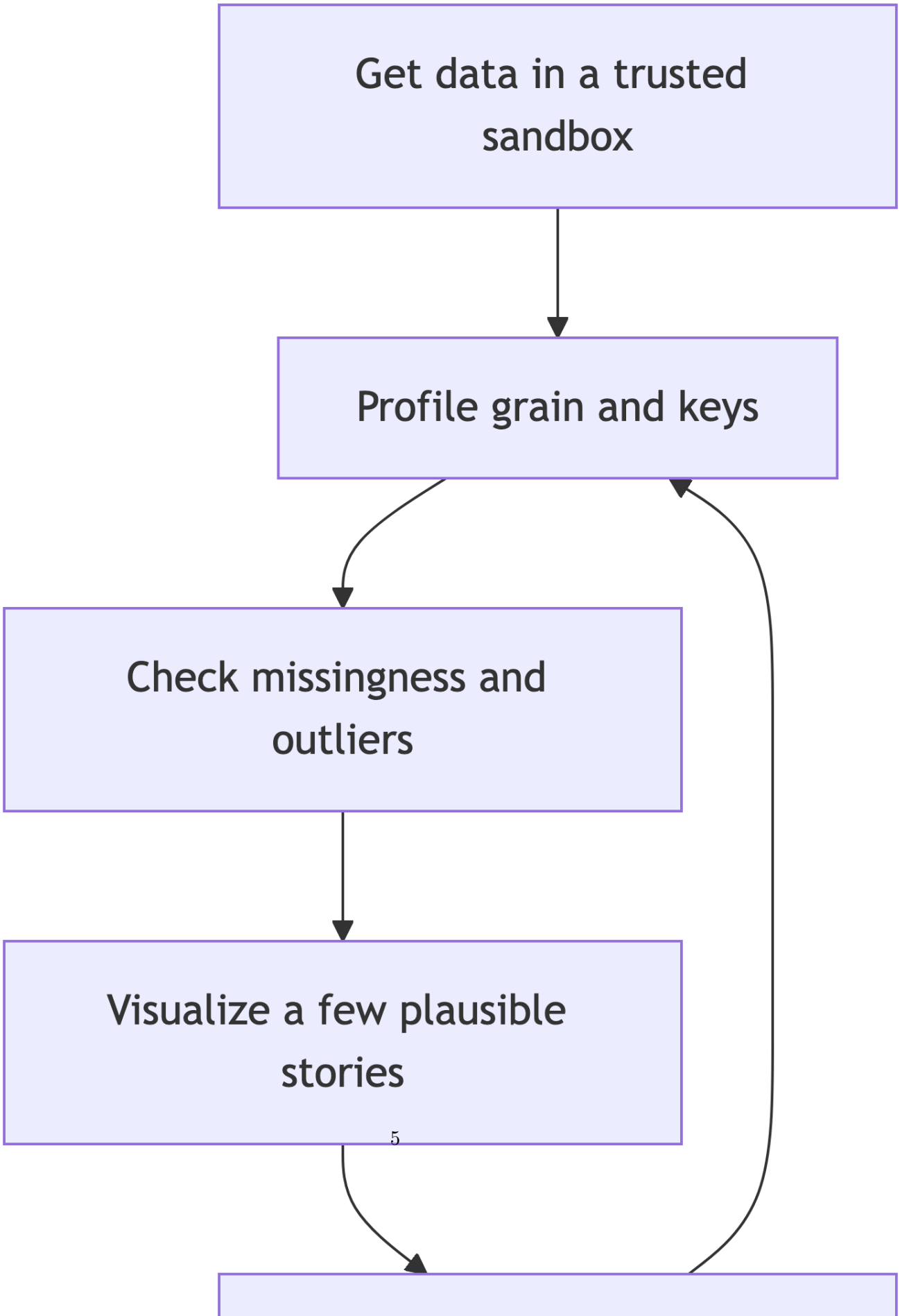
```
A. A single polished chart for the poster background.
B. A short list of data-quality issues with severity and next checks.
C. Twelve significance stars from pairwise t-tests on every column.
D. A trained model with no written definition of the target.
```

. . .

B. Artifacts that change what you measure, split, or fix are the point. C and D optimize the wrong scoreboard; A alone skips the hard lessons.

3.4 Common EDA moves (starter menu)

- **Inventory:** row counts, duplicates, primary keys, slowly changing IDs.
- **Missingness:** MCAR vs MAR vs MNAR is hard, but patterns still matter.
- **Marginals:** histograms, bar charts, value ranges.
- **Relationships:** scatter plots, stratified summaries, simple correlations (read carefully).


Get data in a trusted sandbox
Profile grain and keys
Check missingness and outliers
Visualize a few plausible stories

Encourage them to log findings in the repo (README, issue list, or lab notes) so meta-project peers can follow the trail.

4 Part 3: Research questions

4.1 What is a research question?

4.2 Focused inquiry

A research question is a **focused inquiry** that guides analysis:

- It **articulates** the problem you want to solve or the knowledge you want to gain.
- It is **grounded** in theory, prior studies, or observed phenomena (including EDA surprises).
- It should be **specific enough** that someone else can tell whether you answered it.

. . .

If your question is "understand customers," that is a mood, not a question. Push for outcome, population, comparison, and time window.

4.3 Why research questions matter

They:

- **Provide direction and purpose** so the backlog does not sprawl forever.
- **Define scope** so you can say "in" and "out" for the term.
- **Enable testable hypotheses** so you know what evidence would change your mind.

. . .

Weak questions produce weak reviews from faculty and industry. Strong questions make bad data news early, which is cheaper than bad news at the poster session.

4.4 Types of research questions (taxonomy)

Type	Core move	Example (tightened)
Descriptive	Summarize a phenomenon	What is the distribution of capstone project hours logged per week in DATA 510?
Exploratory	Map structure when theory is thin	Which usage features cluster together among users who later churn?

Type	Core move	Example (tightened)
Inferential	Learn about a population from a sample	Does mean model error differ between two deployed algorithms after controlling for traffic seasonality?
Predictive	Forecast or classify future cases	Can we predict churn in the next 30 days from usage in the prior 90 days?
Causal	Support a claim about causes	Did a policy change *cause* a shift in completion rates, or did seasonality confound it?

. . .

Exploratory questions are still research-shaped: they narrow *what you will look at next*. They are not a license to run every test in the menu.

4.5 Quick Quiz: Classify the question

"We want to know whether our new recommender increases click-through *because* of the model and not because we also changed the UI the same week."

```
A. Descriptive
B. Predictive
C. Causal
D. Purely EDA with no question
```

. . .

C. You are asking about a cause-and-effect claim; that demands design, not just a leaderboard metric. (Also: expect confounding and plan for it.)

5 Part 4: Putting EDA and research together

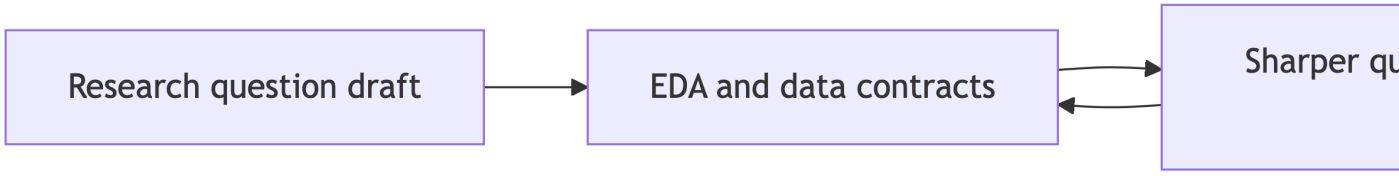
5.1 Research plus EDA

5.2 A loop, not a ladder

Research questions often **lead to** EDA (you need to operationalize constructs and check feasibility).

EDA often **refines or generates** research questions (the data talks back).

. . .



5.3 Complementary strengths

- **Research** supplies focus: what would count as success this term?
- **EDA** supplies ground truth about messiness: what will blow up the plan if ignored?

. . .

Together they support **data-informed** decisions instead of **data-stamped** decisions (pretty charts pasted under a foregone conclusion).

5.4 Guardrail: exploratory work is not p-hacking

If every plot becomes a hypothesis test on the same sample, you are not doing EDA plus research. You are doing **multiple chances to get lucky**.

. . .

Capstone habit: separate **exploratory notebooks** from **confirmatory** analysis paths, pre-register the latter when possible, and document what changed after EDA (even if it hurts).

Optional reading to cite live: Gelman and Loken on researcher degrees of freedom; any short ASA guide on p-values and multiplicity.

6 Part 5: Interactive work

6.1 Try It: From churn question to plan

Predictive research question: Can we predict customer churn in the next 30 days from usage patterns in the prior 90 days?

With neighbors (about 5 minutes total):

1. Name **three EDA moves** you would run first and what each could invalidate.
2. Rewrite the question **one notch sharper** (population, label, horizon, leakage risks).
3. Sketch **one** evaluation approach you would defend to a skeptical mentor.

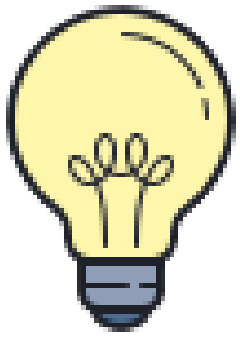


Figure 1: Brainstorm icon

6.2 Quick Quiz: After EDA

You discover the churn label is missing for the entire newest month of data. What is the best next step for the *research* side of the project?

```
A. Drop the month silently so metrics look cleaner.
B. Keep modeling; missing labels are a deployment detail.
C. Document the issue, revise timelines or scope, and agree on a label policy with stakeholde
D. Switch to causal claims because prediction failed.
```

. . .

C. EDA exists partly to force these conversations early. A and B are how projects lose trust; D is the wrong genre swap.

7 Wrap-up

7.1 Big ideas

1. **EDA is for contact with reality:** schema, quality, and shape before heroics.
2. **Research is for commitment:** what you will argue with evidence by the end of the term.
3. **Exploratory vs confirmatory** is a workflow distinction, not an excuse to test everything.
4. **Taxonomy language** helps you communicate with mentors, peers, and meta-project partners.
5. **Loops beat ladders:** expect your question to change; document why.

7.2 References

1. John W. Tukey, *Exploratory Data Analysis* (classic framing of EDA as discovery).
2. Andrew Gelman and Eric Loken, "The garden of forking paths" (2014) on researcher degrees of freedom and exploratory vs confirmatory analysis.
3. American Statistical Association statements on p-values and statistical inference (short, readable primers).

Questions? Pull me aside after block one or use office hours on the syllabus.