Final Project: Learning from Data You Collect

STA 119FS: Statistics as a Way of Thinking | Fall 2026

November 2, 2026

The project

Your group will investigate a question about campus life by collecting and analyzing your own data. You may conduct a survey, run a small experiment, make physical measurements, or systematically observe something happening on campus. You will design the investigation together, carry it out, examine the evidence, and present what you learned. **Each person will write and submit their own final report.**

The central challenge is to decide what observations could answer your question and then judge how well the observations you actually collected do that job. A successful project explains the decisions behind the data, checks the analysis, and makes a claim that the evidence can support. Unexpected findings, weak patterns, and inconclusive results can all support excellent projects.

The final project is **30% of your course grade**. The project itself is graded out of 100 points. You do not need a statistically significant result, a randomized experiment, expensive equipment, or methods beyond those taught in class.

Five milestones and deadlines

All written submissions are due at **11:59 p.m. Eastern time** on the listed date. Presentations will take place during the regular class meeting; slides are due at **9:00 a.m. Eastern time that day**.

Milestone	Deadline, 2026	Submission	Points
1. Two proposed project ideas	Saturday, October 31	One packet per group, covering both ideas	10
2. Plan of action and timeline	Monday, November 9	One per group	15
3. Account of collection and analysis	Monday, November 23	One per group	20
4. Presentation	Thursday, December 3, in class	Group presentation; every person participates	15
5. Final written report	Thursday, December 10	One per person; shared files once per group	40
Total			**100**

The first three milestones receive shared group scores. The presentation includes 10 shared points and 5 individual points. Your report receives an individual score: **55 project points are shared and 45 are individual.** A project score of 90/100 contributes 27 of the 30 available percentage points to your course grade.

Working together

Groups have been formed; find yours in the project module on Canvas. Every member should contribute to design, collection, and analysis or checking, and should understand the whole investigation. Assign named responsibilities and keep a brief contribution log throughout the project, including any changes to those responsibilities.

Choosing a question and collecting useful data

Use the accompanying Campus Data Collection companion to identify your two options. You may propose **two ideas from the bank, or one bank idea and one original idea**. A bank idea may be adapted to campus, but identify the entry and explain the changes. If an adaptation changes the central question substantially, ask the instructors whether it counts as your original idea. The bank contains starting points, not complete protocols or results you are expected to reproduce.

Before you start writing Milestone 1, choose your two options and record your group name and ranked options on the **project sign-up sheet**, linked from the project module on Canvas. Check other groups' entries first: **do not choose a project another group has already registered.** If interests overlap, discuss a distinct question or approach with the instructors; signing up does not itself approve a project.

Each option should have one main question that can be investigated with modest resources and the time available. It must involve active campus collection by your group. Existing datasets and research may provide background, but cannot replace your own collection. These are just short descriptions of six of the examples. Full descriptions and other options are available in the companion file.

Starting interest	A question you could investigate	A measurement issue to resolve
Paper-airplane reward choices	How do people choose between throwing distance, difficulty, and reward?	Distance can reflect skill, confidence, or willingness to accept risk.
Blind water tasting	Which water sample do participants prefer, and can they distinguish the samples?	Preference and correct identification are different outcomes.
Crowd estimation	How do mean and median guesses compare with a known quantity?	Repeated guesses from one person are related observations.
Campus preferences	How do respondents' estimates of a preference's popularity compare with the responses you collect?	Your sample may not describe all Duke students.
Door holding	How often is a door held at different follower distances?	Define an eligible opportunity and a consistent distance measure.
Library seating	How does the fraction of usable seats occupied differ across selected times?	Does a seat claimed by belongings count as occupied or available?

Requirements for every project

1. **Connect the question to the measurement.** Define the population or setting you hope to learn about, what you can actually observe, and the gap between those two. Explain what one row of data represents. Record units and define categories so another group member could use them consistently.
2. **Plan who, where, and when.** State how observations enter your dataset, who or what is left out, and what comparison you will make. Distinguish a convenience sample from random sampling, and random sampling from random assignment. An observational comparison can be valuable without establishing cause and effect.
3. **Pilot the procedure.** After the instructors approve the idea and pilot activity, try a small rehearsal before main collection. Check the wording, timing, recording form, equipment, and consistency between observers. Mark pilot records separately and explain whether they will be included in the analysis.
4. **Choose a defensible scope.** Propose a collection target and a rule for when to stop, based on feasibility and the variation you need to observe. There is no universal minimum sample size. Forty people making three guesses produce 120 guesses, not 120 independent people. Do not keep collecting solely until a preferred result appears.
5. **Keep a record of what happened.** Preserve original observations, log collection sessions and departures from the plan, and document corrections or exclusions. Retain valid observations even when they are surprising. Explain any shortfall relative to your target.
6. **Treat people and spaces respectfully.** Participation in surveys or activities must be voluntary, with a brief explanation and an easy way to decline. Collect only information needed for the question; use anonymous labels rather than names or contact details. For passive observation, record events without identifying people. Discuss site access, food, rewards, or other special arrangements with the instructors before starting. Keep non-identifying participant records in your group's assigned private GitHub repository.

Class project and participant notice. These projects are educational activities for STA 119FS, intended to develop your skills rather than contribute to a field of knowledge. Data and results will stay within the course and will not be made public. Under Duke's IRB Review of Student Research and Class Projects policy, projects meeting these conditions do not constitute research with human subjects and do not require IRB approval. Instructor approval is still required before piloting or collecting data.

Supplies, budgets, and rewards

Each group will have the same fixed supply budget of **$50**. Anything left over will go toward a larger donut haul in December. Include an itemized budget with Milestone 2, even if your project costs nothing. List each item, quantity, unit price, total cost, and a product link or other basis for the estimate. Separate items you can borrow from those you need purchased. Include any nonmonetary rewards in the total.

Submit purchase requests to the instructors before spending; the proposed arrangement is for instructors to buy approved supplies so students do not need to pay out of pocket. Obtain approval before exceeding the group cap. Do not assume reimbursement for unapproved purchases. If your pilot needs purchased supplies before November 9, request them with the pilot approval.

Rewards must be nonmonetary. Points, stickers, or small approved items are possible; cash and cash-equivalent rewards are not. State in advance whether a reward is for participating or for a particular outcome, and describe the exact rule. For an airplane task, for example, specify what counts as hitting the target, how points vary with distance, what a miss earns, and whether any partial credit is possible. The companion gives examples to adapt.

From the plan to the final report

The following is a suggested working sequence. Adjust it in your plan while preserving the five milestone deadlines.

Working period	Main task	Evidence of completion
October 15-31	Choose two options and register them on the sign-up sheet, then develop both proposals	Milestone 1
November 1-9	Meet with the instructors, select one project after feedback, run its approved pilot, and revise the protocol and budget	Milestone 2 and its appendices
November 10-18, after plan approval	Collect main data and check recording quality	Collection log and original records
November 19-23	Clean data, analyze, and check results	Milestone 3
November 30-December 2	Address analysis feedback and rehearse	Slides and speaking plan
December 3	Present, discuss the evidence, and record feedback	Milestone 4
December 4-10	Respond to feedback and finish individual reports	Milestone 5 and final shared files

Looking at the data honestly

Keep the dated action plan as your record of decisions made before main collection. You may correct a problem or investigate a new pattern later. Record what changed, when, and why, and distinguish the original analysis from later exploration. Do not rewrite the original question as though you had predicted an unexpected result.

Use the main summary you planned even if another one tells a more exciting story. Explain any valid reason to depart from the plan. Report relevant missing observations, unsuccessful attempts, and exclusions; do not quietly remove inconvenient cases. Check unusual values against the records before deciding whether they are errors.

Use course-level R for summaries and figures, with readable code or documented steps that reproduce the reported numbers. Descriptive analysis may fully answer your question. Formal tests, confidence intervals, regression, or simulations are optional when appropriate and understood. If you use them, explain the assumptions and account for repeated observations. A calculation cannot by itself remove selection or measurement problems.

Milestone 1: Two proposed project ideas

Group submission | 10 points | Due date: October 31

There are two things to show us at this first check-in:

1. **Your first-choice project proposal.**
2. **Your second-choice project proposal, developed as a workable alternative.**

At least one must be based on a project-bank entry; at most one may be original. Submit a single group packet with **1-2 pages per idea (2-4 pages total)**, plus references if needed. Put all group members' names on the packet. Label each proposal with its rank, title, and bank entry number or "original idea." Add a brief explanation of your ranking. The two options must match the ones your group registered on the **project sign-up sheet** (linked from the project module on Canvas) before writing.

Address all six items below for each proposal. A title or a one-sentence backup idea is not enough. We should be able to judge whether either project is interesting, answerable, and feasible.

1. **Question and motivation.** State one main question in a sentence. Why is it worth investigating? Identify the idea-menu entry or other source that inspired you, if applicable, and explain how your version differs.
2. **Possible answers.** Describe what different results might look like. If you have a prediction, give its reasoning. Say what would surprise you. A directional prediction is optional.
3. **Data you would need.** Identify the population or setting of interest, the proposed unit of observation, the main outcome, and any groups or conditions to compare. Give concrete examples of what you will record, including units or response options.
4. **Collection approach.** Describe the proposed campus locations, recruitment or observation method, approximate scale, supplies, rough costs, and access arrangements. Explain why the work can be completed on time. The full itemized budget comes with Milestone 2.
5. **Likely difficulty.** Identify two specific ways the measurements or sample could mislead you. Explain one claim your proposed study would not be able to support.
6. **Next step.** Outline a small pilot and identify the most important question you want instructor feedback to resolve.

The two proposals are alternatives: **you will carry out only one approved project**. After this check-in, your group will meet with the instructors to discuss both proposals; we will help you select and refine one of them. You do not need collected data at Milestone 1, and should obtain approval before piloting either idea. A clear explanation of an unresolved problem is more useful than pretending every design choice has been settled.

Milestone 2: The plan of action and timeline

Group submission | 15 points | Due date: November 9

Submit **one PDF containing a 2-3 page plan for the one project selected after feedback, followed by appendices**. The 2-3 page limit applies to the main plan only. Explain which idea you selected and why, incorporate instructor feedback, and report what you learned from the approved pilot. Address the seven components below in the main plan, placing the complete collection instrument and participant notice, itemized budget, draft data dictionary, and other detailed supporting materials in the appendices. The plan should be specific enough that another student could follow it.

Participant notice for the pilot and main collection. Include the exact notice below in your collection protocol and explain when and how you will display or read it:

> We are students completing a class project for STA 119FS at Duke. Participation is voluntary; you may decline or stop at any time. The data are for class use only and will not be made public. Please do not provide your name or contact information.

Use this notice during the pilot. In your plan, explain any changes after the pilot to when or how you provide it. **Use this participant notice whenever recruiting or interacting with participants during main data collection.** Explain how your records will prevent identification, including through combinations of details. For passive observation, explain how you will provide notice where appropriate; consult the instructors if you are unsure how to handle your setting.

Include these seven components:

1. **Final question and collection protocol.** Provide the actual survey questions and response options, observation sheet and category definitions, or experiment instructions and assignment procedure. Define eligibility, the unit of observation, main measurements, and any denominator. If randomizing, explain how; if observing, explain how eligible events are selected.
2. **Sampling and stopping.** Set the target number of people, events, locations, or sessions as appropriate; explain why it is workable. List collection locations and time windows, how repeat participation or repeat observations will be handled, and the stopping rule. Identify likely omissions and what you can do about them.
3. **Pilot findings and revisions.** Describe what you tried, what failed or was unclear, and what changed. If you retain the procedure, explain what the pilot checked and why no revision was needed. State how pilot data will be handled.
4. **Analysis choices made in advance.** Specify the main outcome, comparison, summary, and plot. State initial rules for missing values, incomplete attempts, duplicate records, and exclusions. Plan at least one meaningful sensitivity check: repeat an analysis using another reasonable choice and see whether the conclusion changes. For example, compare mean and median guesses, or count belongings-claimed seats in two defensible ways. Explain why your check matters for your question.
5. **Responsibilities and calendar.** Give dated tasks, a named lead and checker for each, and the files or results each task will produce. Include a backup for low recruitment, weather, equipment failure, or unavailable space. Everyone should do substantive work and check work beyond their own assigned piece.
6. **Data organization.** Provide a draft data dictionary explaining each variable, its units or categories, and how missing values are represented. Explain how group members will organize files in the assigned private GitHub repository and use consistent versions.

7. **Itemized budget and reward rules.** Provide quantities, unit prices, totals, and product links or other price estimates for requested supplies and nonmonetary rewards. Note borrowed items, confirm whether the total fits the announced cap, and identify any additional funding request. If no purchases or rewards are needed, say so. Explain any scoring or reward rule in the collection protocol and request instructor approval before purchases.

Required private GitHub repository

The instructors have created a **PRIVATE GitHub repository** for each group in the course organization. **You are required to use your assigned repository for all project files and submissions from Milestone 2 onward.**

By Milestone 2, upload:

- **Your group's Milestone 1 submission:** a copy of the packet you already submitted.
- **Your Milestone 2 plan PDF:** the 2-3 page plan and all appendices together in one file. The collection instrument, participant notice, budget, and draft data dictionary belong in this PDF; they do not need separate PDF uploads.
- **Data and working files, separately:** any non-identifying pilot data, collection or decision logs, and code you have created. Upload data in usable files such as CSVs or spreadsheets, and code as scripts or analysis documents. Also include any editable forms or spreadsheets you will use for collection or recordkeeping, even if a copy appears in the appendices.

As the project progresses, keep all project files in this repository, including non-identifying original and cleaned data, code, figures, tables, logs, slides, and later milestone submissions. Each student must upload their own final report as a separate PDF labeled with their name.

Milestone 3: How collection and analysis went

Group submission | 20 points | Due date: November 23

Submit a **3-4 page account**, including preliminary tables and figures, plus the working data and analysis files. This should be a substantive account of an analysis you have carried out. Include the main planned result and a completed sensitivity check; identify what still needs work before the presentation and final reports.

Organize the account around these questions:

1. **What did you actually collect?** Compare the plan with the achieved sample, locations, dates, and conditions. Report relevant counts: approached or eligible, participated or observed, completed, excluded, and analyzed. If a count could not be observed, say so. Separate unique people or units from repeated records. Give the denominator for each rate or percentage.
2. **What did the data look like before analysis?** Describe missing values, ambiguous responses, recording mistakes, unusual values, uneven collection, or disagreement between observers. Explain which problems you corrected, which remain, and which might affect your conclusions. If checks revealed no problems, report the checks and their results.
3. **How did you get from original records to results?** Explain coding, cleaning, exclusions, and derived variables. Record the number of observations affected. Include a short decision log with the date, decision, reason, and whether the decision came before or after examining outcomes.
4. **What is your best current answer?** Show at least **one summary table and two purposeful figures**. Label axes, units, groups, and sample sizes; explain in words what each display contributes. Present the main planned comparison and clearly label any exploratory analysis. You may combine related panels into one figure if each has a clear purpose.
5. **What happened when you checked the answer?** Complete at least one sensitivity check and compare the results directly. Report the relevant numbers or plots under both choices. Explain whether the conclusion changes and what that tells you about the measurement or analysis.
6. **What surprised you, and what remains uncertain?** Compare findings with your initial reasoning. A result can reflect a real pattern, chance variation, selection, measurement, or several of these together. Discuss plausible explanations without claiming to have established one. If nothing was surprising, explain why the findings fit your expectations and what remains unresolved.
7. **What will you do before the presentation and final report?** List a few concrete remaining tasks with owners and dates. These might include checking a plot, resolving a coding discrepancy, or narrowing a claim. Additional collection requires an explicit reason and a documented plan; do not extend collection just to obtain a preferred result.

Confirm how you provided the participant notice during main collection and how you kept participants from being identified. Explain any departures from the approved procedure and how you addressed them with the instructors.

Working files to include

Provide the current cleaned dataset, data dictionary, R script or analysis document, collection log, and cleaning/decision log. Include the original non-identifying records in your group's assigned private GitHub repository along with the account and all working files. Files may be simple: a short script and a clearly labeled spreadsheet can be sufficient. The goal is that a reader can follow what happened and reproduce the main numbers.

Feedback on this milestone should help you improve the presentation and final reports. Finding and explaining a genuine problem is evidence of good judgment. What matters is how you investigate it and adjust your conclusion.

Milestone 4: The presentation

Group presentation | 15 points: 10 shared, 5 individual | Date: December 3

Plan an **XX-minute presentation** (exact timing to be announced when we finalize whether we will use 1 or 2 days for presentations), **followed by 4 minutes of questions**. Every student must make a substantive spoken contribution and be prepared to explain the investigation. An introduction or slide transition alone is not a substantive contribution. Arrange any needed presentation accommodations with the instructors.

Give the audience a clear account of the question, what you did, what the evidence says, and how confident they should be in the conclusion. Include:

- The question and why the observation or activity was a useful way to investigate it.
- Enough detail about collection and measurement to understand the evidence.
- Two or three readable displays of the main results, including the key sensitivity comparison when useful.
- One consequential difficulty, surprise, or limitation and how you dealt with it.
- A connection to a course idea, a defensible conclusion, and a concrete next step.
- How someone from around 1700 or 1800 would have analyzed your data differently. Explain why they would have done it differently and whether you think they would have gotten a different answer. You do not have to carry out that analysis.

For a project based on the bank, include a comparison with its original study in both your slides and your spoken presentation. State how your question or procedure differs, compare findings where results are available, and discuss plausible reasons for similarities or differences: participants, campus setting, timing, measurement, incentives, sampling, or chance. If you cannot access the original results, say so and discuss what could make two studies on different campuses produce different results. Do not invent a numerical comparison or treat the original finding as a target to match. Cite the source you actually used.

The presentation should contain a complete analysis and a defensible conclusion. Record questions and comments from the discussion to inform your individual reports.

Use readable displays and explain axes and denominators aloud. Rehearse and leave time for questions. Every student should be able to explain a main figure and a limitation, including material outside their own segment. Questions assess reasoning; recognizing what the data cannot answer can be a strong response. Instructors will direct a question to anyone who has not yet answered to assess individual understanding.

Milestone 5: Your individual final report

Individual submission | 40 points | Due date: December 10

Each student submits a **1,800-2,400 word report**, excluding figures, tables, references, and the contribution statement. Use informative headings and write for a classmate who has not worked on your project. The report must stand on its own: a reader should understand the question, data, analysis, and conclusion without opening a teammate's report.

Use the week after the presentation to consider feedback and refine your account. Additional data collection is not required. Your report should contain:

1. **The question and a brief answer.** Introduce the question and its motivation, then give the main finding in a short opening summary. Indicate the population or setting your answer actually describes.
2. **Design and measurement.** Explain recruitment or observation, locations and timing, the unit of observation, what was measured, and the intended comparison. Describe the pilot and important departures from the plan. Make clear whether the design supports a descriptive, associational, or causal claim.
3. **The data you ended up with.** Give final sample counts and denominators, including repeated measurements, missingness, and exclusions. Evaluate the most consequential data-quality problems and explain how they could change the answer. Be specific about this dataset.
4. **Analysis and evidence.** Present the main analysis, at least one summary table and two purposeful figures, and the sensitivity check. Explain magnitudes in meaningful units, not merely whether numbers differ. Distinguish planned analyses from exploratory ones. Discuss variability and what the observed evidence does and does not resolve.
5. **Interpretation.** Answer the original question as far as the data allow. Address unexpected, weak, or conflicting findings without explaining them away. Consider alternative explanations and limits on generalization. No clear observed difference is not proof that two things are identical; a correlation alone does not establish a causal effect.
6. **Connections to the course.** Develop at least **two specific connections** to ideas from readings, lectures, or labs. At least one should explain a historical problem or approach encountered in the course and how it clarifies a decision in your project. Possible connections include combining discordant measurements, defining an average, choosing denominators, distinguishing measurement error from variation, or interpreting regression and correlation. Explain the connection rather than listing names or methods. Cite the relevant reading, lecture, or lab.
7. **Feedback, improvements, and learning.** Address at least one substantive question or comment from the presentation. Explain what you revised in response, or why you retained your analysis or interpretation. If no substantive feedback was received, address an unresolved question you identified during the discussion. Propose one or two concrete improvements to a follow-up investigation and connect each to an observed limitation. End with what you learned.

Include a **100-150 word individual contribution statement** after the report. Refer to the responsibilities in your action plan: identify assigned tasks you completed, assigned tasks you did not complete (and why, including any agreed reassignment), and work you completed that was not originally assigned to you. If there are no items in a category, say so. Include what you checked and one decision you would make differently. Add references and the required acknowledgments of shared work, sources, code, and tools. Appendices may document details, but the main report must contain the evidence needed to assess your conclusion.

Submission and collaboration details

From Milestone 2 onward, submit all project work through your group's assigned PRIVATE GitHub repository in the course organization. Upload milestone PDFs and every associated file by the listed deadlines. Also upload a copy of your Milestone 1 submission by Milestone 2. For group submissions, designate one uploader and put every member's name on the submission. Each student uploads their own final report as a separate PDF labeled with their name. Upload presentation slides as a PDF to the same repository by 9:00 a.m. on presentation day. Check that submitted files are visible on GitHub; saving them only on your computer does not submit them.

Keep project data, results, slides, and reports in private course spaces. Do not post them in public repositories, on websites, or on social media.

For Milestone 3, upload the account and all required supplementary files to clearly labeled folders in the repository.

At Milestone 5, organize the final shared files in a clearly labeled folder in the repository containing the non-identifying original records, cleaned data, data dictionary, final code, collection and decision logs, and a brief README explaining which file to run to reproduce the main results. Include the final shared figures and tables. Use relative file paths where practical. Check that the code works from the submitted files and that final numbers agree across the report, figures, and shared files. Preserve the submitted slides and earlier file versions. Date any changes after the presentation and explain how they affect the final results; do not replace earlier findings.

Page and word targets guide scope; prioritize clarity and completeness. Put instruments and detailed logs in appendices. Advanced programming and elaborate document design are not required.

Raise unequal workloads or missed responsibilities with your group early; contact an instructor if the problem persists. The contribution log and individual statement document specific work and help resolve concerns. The grading scheme uses shared scores for shared products and individual scores for the report and individual presentation criteria.

The syllabus's academic-integrity and AI rules apply. AI may support learning and code debugging after your own attempt; verify and understand the code. Do not paste the assignment into AI or ask it to complete the work. Submitted narrative cannot include AI-generated prose, even after editing. Disclose material AI assistance with the tool, model, date, and complete relevant transcript link. Cite borrowed ideas, language, and code.

The syllabus's two late allowances apply to labs and homework. **Project extensions require an arrangement with an instructor.** Contact the instructors promptly if a problem affects a milestone or the group's collection schedule. Approved accommodations and university-authorized absences continue to apply.

Shared material and independent authorship

Group members may share the dataset, code, numerical results, tables, figures, and discussion of the findings. Identify shared material as group work. **Write all final narrative and figure captions independently.** Do not divide the final report into sections to be written by different people, circulate report drafts for teammates to rewrite, or copy and lightly rephrase the group's earlier prose. Explain the shared methods and findings in your own words. Your interpretations may reasonably differ; you do not need to manufacture a different result or run an extra analysis to make your report individual.

Final submission checklist

- The question, unit of observation, sample counts, and denominators are clear.
- The account distinguishes the original plan, actual collection, and later exploration.
- The report contains a summary table, two purposeful figures, and an interpreted sensitivity check.
- The conclusion addresses uncertainty and the specific limitations of the data.
- Two course connections are developed, including a historical connection.
- A bank-based project's slides and talk compare with its original study, or discuss possible across-campus differences when original results are inaccessible.
- Each person has written their own report and captions and accounted for completed, uncompleted, and additional work in a contribution statement.
- Sources, shared material, and permitted tool assistance are acknowledged.
- All milestone submissions, including a copy of Milestone 1, individual final reports, slides, data, dictionary, code, logs, displays, and README are in the assigned private GitHub repository.
- Each student presented substantively and addressed feedback in their report.