STAT 204 Final Project Guidelines

Winter 2025

Table of contents

1 Overview **2**

2 Project Tracks **2**
- 2.1 Track 1: PhD Students in Statistics (Individual Work) . . . 3
- 2.2 Track 2: Students from other programs (Pairs) . . . 3

3 Project Proposal (Week 6) **3**
- 3.1 Written Proposal (½ - 1 page) . . . 3
- 3.2 Proposal Presentation (5 minutes) . . . 4

4 Final Presentation (Week 10) **4**
- 4.1 Presentation Structure . . . 4
  - 4.1.1 For All Tracks: . . . 4
- 4.2 Additional Requirements: . . . 5

5 Final Report (Due December 9, 2025 by 5 PM) **5**
- 5.1 Report Structure . . . 5
  - 5.1.1 1. Introduction (0.75-1 page) . . . 5
  - 5.1.2 2. Exploratory Data Analysis (1-1.5 pages) . . . 5
  - 5.1.3 3. Methods (1-1.5 pages) . . . 6
  - 5.1.4 4. Results (2-2.5 pages) . . . 6
  - 5.1.5 5. Discussion & Conclusion (0.75-1 page) . . . 7
  - 5.1.6 6. References . . . 7
- 5.2 Writing Expectations . . . 7
- 5.3 Figures and Tables . . . 8

6 Track-Specific Expectations **8**
- 6.1 Track 1: PhD Statistics Students . . . 8
- 6.2 Track 2: Student from Other Programs . . . 9

7 Methods You Might Use **9**

8 Assessment Rubric **10**
8.1 Proposal (5% of final project grade) 10
8.2 Presentation (30% of final project grade) 10
8.3 Final Report (65% of final project grade) 10

9 Tips for Success **10**

10 Submission Details **11**
10.1 Proposal: 11
10.2 Presentation: 11
10.3 Final Report: 11

11 Getting Help **12**

12 Questions? **12**

1 Overview

The final project is your opportunity to demonstrate the data analysis skills you've developed throughout this course. You will conduct a complete statistical analysis—from initial data exploration through model building, diagnostics, and interpretation—and communicate your findings through written and oral presentation.

Timeline:

- **Week 6**: Project proposal due + 5-minute presentation during lecture
- **Week 10**: Final presentations during lecture time (10 minutes each)
- **December 9, 2025 by 5 PM**: Final report due

---

2 Project Tracks

Students will complete different versions of the project based on their program and background:

2.1 Track 1: PhD Students in Statistics (Individual Work)

Focus: Paper reproduction, critical analysis, and methodological extension

You will select a published paper that uses methods covered in this course, reproduce the main analysis, critically evaluate the approach, and propose an extension or alternative model.

2.2 Track 2: Students from other programs (Pairs)

Focus: Data-driven research question and comprehensive analysis

You will identify a dataset and research question of interest, conduct exploratory and/or confirmatory analysis using course methods, and interpret findings in context. If you want, you can choose to complete Track 1 instead.

---

3 Project Proposal (Week 6)

3.1 Written Proposal (½ - 1 page)

Your proposal should include:

1. **Project identification**
    - Track 1: Full citation of the paper you will reproduce
    - Tracks 2: Description of your dataset (source, size, variables)
2. **Research questions or objectives** (2-4 specific questions/goals)
3. **Preliminary methods plan**
    - Which course topics/methods do you anticipate using? (e.g., linear regression, logistic regression, PCA, clustering, regularization)
    - Why are these methods appropriate for your questions?
4. **Data access confirmation**
    - Track 1: Confirm you have access to the data or code from the paper
    - Track 2: Confirm you can access and load the dataset in R
5. **Team member names and roles** (for paired projects)

3.2 Proposal Presentation (5 minutes)

Presentations will be given during lecture time in Week 6.

Prepare slides covering:

- Your research question or paper of interest
- Why this topic matters
- Brief overview of your dataset or the original study
- Planned analytical approach

Be prepared for questions and feedback from classmates and instructor.

---

4 Final Presentation (Week 10)

Format: 10 minutes + 2-3 minutes for questions

Presentations will be given during lecture time in Week 10.

4.1 Presentation Structure

4.1.1 For All Tracks:

1. **Introduction (1-2 min)**
    - Research question or paper objective
    - Why this analysis matters
    - Dataset overview
2. **Methods (2-3 min)**
    - Analytical approach
    - Key models/techniques used
    - Brief justification for method choices
    - Track 1: Critical evaluation and proposed extension
3. **Results (4-5 min)**
    - 3-5 key findings with visualizations
    - Model diagnostics (where relevant)
    - Interpretation of main effects/patterns
    - Track 1: Results of proposed extension

4. **Conclusion (1-2 min)**
    - Answer to research questions and limitations
    - Future directions

4.2 Additional Requirements:

- Include **at least 2 high-quality visualizations** (exploratory or model-based)
- Show evidence of **model diagnostics** for at least one fitted model
- Keep technical jargon accessible—your audience includes students from various backgrounds
- **Peer feedback**: You will complete brief feedback forms for other presentations (attendance mandatory)

---

5 Final Report (Due December 9, 2025 by 5 PM)

Page limit: Maximum 6 pages (including figures and tables, excluding references) **Format**: Must be done using Quarto, you can choose one of the templates for journal articles listed in here: https://quarto.org/docs/journals/

5.1 Report Structure

5.1.1 1. Introduction (0.75-1 page)

- Research question or paper overview
- Background and motivation (make sure to include proper citations)
- Dataset description (source, size, key variables)
- Study objectives

5.1.2 2. Exploratory Data Analysis (1-1.5 pages)

- Summary statistics
- Data visualization (2-3 key plots)
- Data quality assessment (missing values, outliers, distributions)
- Preliminary insights that inform modeling choices

Track 1 (Paper Reproduction) should include:

- A combination of original EDA and useful pieces of the EDA presented in the paper. If you use/reproduce a plot from the original paper in your report, give credit to the authors in the source.

5.1.3 3. Methods (1-1.5 pages)

Track 1 (Paper Reproduction) should include:

- Short explanation of the methods used in the original paper
- Proposed extension or alternative model (following the next section details)

Both tracks should include:

- **Statistical models used** with mathematical notation where appropriate
  - For regression: specify model equation, link function (if applicable), assumptions
  - For PCA: explain dimensionality reduction approach
  - For clustering: describe method and distance metric
  - For regularization: specify penalty type and selection procedure
- **Why these methods?** Connect to research questions and data structure
- Software and key R packages used (don't forget to cite the R packages)

5.1.4 4. Results (2-2.5 pages)

Track 1 (Paper Reproduction) should include:

- Summary of reproduction success/challenges
- What did the original authors do well? What could be improved?
- Summary of results with proposed model following the next section details.

Both tracks should include:

- Model fitting and diagnostics
  - Assumption checking (residual plots, normality tests, etc.)
  - Model comparison (if applicable)
  - Goodness-of-fit measures
- Parameter interpretation with confidence intervals where appropriate
- Key findings presented with visualizations

5.1.5 5. Discussion & Conclusion (0.75-1 page)

Both tracks should include:

- Answers to research questions with supporting evidence
- Practical implications or insights
- Limitations and assumptions
- Future directions

Track 1 (Reproduction) should include:

- A summary of what could be improved in the original paper and how your new analysis does exactly that.

5.1.6 6. References

- Not included in page count
- Use consistent citation style (APA, Chicago, or similar)
- Include at least 2-3 references beyond the dataset source

5.2 Writing Expectations

All students should demonstrate:

- **Clear, concise scientific writing**
- **Correct statistical interpretation**
  - Distinguish between correlation and causation
  - Properly interpret p-values, confidence intervals, and effect sizes
  - Avoid overstating conclusions
- **Reproducibility**
  - Code should be well-commented and organized
  - Include a brief appendix (not counted in page limit) with session info: `sessionInfo()` output
  - If your analysis take too long to include it in your .qmd file, provide a folder with necessary files to run your .qmd file. You can also submit your work as a repository on GitHub.
- **Integration of feedback** from proposal and presentation

5.3 Figures and Tables

- All figures must have descriptive captions and labeled axes
- Tables should be formatted professionally with clear headers
- Referenced in text: "As shown in Figure 1…"
- Use color-blind friendly palettes

---

6 Track-Specific Expectations

6.1 Track 1: PhD Statistics Students

Higher expectations for:

- **Mathematical rigor**: Include key derivations or theoretical justifications
- **Critical analysis**: Deeper evaluation of assumptions, alternative approaches, and methodological limitations
- **Extension proposal**: Well-developed alternative model with preliminary theoretical or simulation-based justification
- **Statistical depth**: More sophisticated model diagnostics and sensitivity analyses

Your critique might address:

- Were the model assumptions justified?
- Were there alternative methods that might be more appropriate?
- How robust are the findings to model specification?
- What additional analyses would strengthen the conclusions?

Your proposed extension should:

- Address a limitation you identified
- Be feasible
- Draw on statistical methods covered in class or related literature

---

6.2 Track 2: Student from Other Programs

Emphasis on:

- **Applied analysis**: Thorough, well-justified analytical workflow
- **Interpretation**: Clear connection between statistical results and domain context
- **Communication**: Accessible explanations of technical content
- **Collaboration**: Both team members should contribute substantially; briefly describe division of labor in report

Your analysis should:

- Use at least 2-3 different methods from the course
- Show meaningful exploratory analysis that motivates modeling choices
- Interpret findings in the context of your research question
- Acknowledge limitations honestly

7 Methods You Might Use

Your project should demonstrate command of methods covered in STAT 204. Consider using:

- **Linear regression**: prediction, inference, diagnostics
- **ANOVA**: comparing groups, interaction effects
- **Logistic regression**: binary outcomes, odds ratios
- **PCA**: dimensionality reduction, data visualization
- **Clustering**: k-means, hierarchical methods
- **Regularization**: ridge, lasso, elastic net for high-dimensional data
- **Model selection**: cross-validation, AIC/BIC
- **Bootstrap or simulation**: for inference or sensitivity analysis (advanced)

You are not required to use all of these—select methods appropriate to your research question and data.

---

8 Assessment Rubric

8.1 Proposal (5% of final project grade)

- Clear research question/objectives
- Appropriate methods identified
- Feasible scope

8.2 Presentation (30% of final project grade)

- Clarity and organization (10%)
- Quality of visualizations and results (10%)
- Time management and delivery (5%)
- Peer feedback participation (5%)

8.3 Final Report (65% of final project grade)

- Introduction and motivation (5%)
- Exploratory data analysis (10%)
- Methods appropriateness and justification (15%)
- Results and interpretation (20%)
- Discussion quality (10%)
- Writing quality and reproducibility (5%)

Track-specific adjustments:

- Track 1: Heavier weight on critical analysis and extension
- Tracks 2 & 3B: Heavier weight on applied interpretation
- All tracks: Correct statistical interpretation is essential

---

9 Tips for Success

1. **Start early**: Data cleaning and exploration take longer than you think
2. **Meet with your partner regularly** (for paired projects): Use version control (Git/GitHub) to collaborate effectively
3. **Don't overcomplicate**: A thorough analysis of a focused question is better than a superficial analysis of many questions

4. **Visualize, visualize, visualize**: Plots should tell a story and support your findings
5. **Check assumptions**: Every model has assumptions—check them and discuss what you found
6. **Interpret, don't just report**: Say what your results *mean*, not just what they are
7. **Seek feedback**: Use office hours before each deadline
8. **Incorporate presentation feedback**: Your report should be improved based on suggestions from Week 10
9. **Proofread**: Typos and unclear writing undermine otherwise good work
10. **Make it reproducible**: Someone should be able to run your code and get your results

---

10 Submission Details

10.1 Proposal:

- Submit via Canvas by start of Week 6 class
- Bring slides on USB or email before class for your lecture presentation

10.2 Presentation:

- Upload slides to Canvas by 9 AM on your presentation day
- Presentations during Week 10 lecture time

10.3 Final Report:

- Submit PDF via Canvas by **5 PM on December 9, 2025**
- Include link to code repository or submit .R/.qmd files separately
- Late submissions: -10% per day

---

11 Getting Help

- **Writing support**: Here's a document with recommendations for writing from Prof. Paul Parker.
- **Collaboration policy**:
  - Paired projects: work together, write together
  - Individual projects (Track 1): you may discuss ideas with others but all work must be your own
  - You may use online resources but must cite them
  - Please DO NOT use genAI for writing code or improving writing. Using grammar correction software is OK, but please don't use any software (LLMs, genAI) for other parts of the process of writing.

---

12 Questions?

Review the syllabus at https://malfaro2.github.io/STAT204/2025/syllabus.html for course policies.

Contact me during office hours or via email if you have questions about project scope, data sources, or expectations.

Good luck! I'm excited to see what you discover in your data.