QM 2023: Semester-Long Capstone Project

Course: Statistics II: Data Analytics
Semester: Spring 2026
Instructor: Dr. Cayman Seagraves

Overview

The capstone project is the centerpiece of QM 2023, designed to simulate a real-world quantitative analysis workflow. You will work as a **Junior Quantitative Analyst** for a real estate investment firm, analyzing REIT market dynamics through data engineering, econometric modeling, and professional communication.

Capstone Milestones: 200 points
Final Presentation: 100 points
Total capstone-related points: 300 points
Format: Team-based (3-4 students) with individual accountability
Duration: Weeks 1-15 (entire semester)

Research Prompt

> *"You have been hired as a Junior Quantitative Analyst at a real estate investment firm. The Investment Committee is concerned about REIT return sensitivity across sectors following the 2022-2023 Federal Reserve interest rate hikes. You must build a data pipeline to analyze historical sensitivity of REIT returns to interest rates and economic factors, and estimate the causal impact of market shocks on different REIT sectors."*

Learning Objectives

By completing this capstone, you will:

1. **Execute an end-to-end data science workflow** from raw data to publication-ready results
2. **Apply econometric methods** (Fixed Effects, DiD, time series) to real financial data
3. **Make and defend methodological choices** using economic reasoning and statistical diagnostics
4. **Collaborate effectively** in teams while maintaining individual accountability
5. **Communicate technical findings** to non-technical audiences (Investment Committee simulation)
6. **Use AI tools responsibly** through the "Disclose, Verify, Critique" framework

---

First Step: Choose Your Dataset

The first step in your capstone project is choosing your dataset. You have two options:

Option 1: REIT Master Dataset (Default)

- **Core Dataset**: REIT Master dataset - Monthly returns and characteristics for 100+ US equity REITs (monthly, REIT-level)
- **Supplementary Data**: You can enhance your analysis by adding other datasets to complement the REIT Master data. Common examples include:
  - **FRED Economic Data**: Interest rates, inflation, unemployment (fetched via API)
  - **Other datasets from Orbis Open Dataset Catalog**: Housing data, labor market indicators, climate data, etc.
  - **Other public data sources**: Any publicly available data that enhances your research question
- **Why choose this?** Consistent with weekly assignments, peer support available, well-documented, and aligned with course examples
- **No approval needed** - this is the default track (REIT Master is required; supplementary data is your choice)

Option 2: Alternative Dataset from Orbis Open Dataset Catalog

- Browse and select datasets from the **Orbis Open Dataset Catalog** at phdai.ai
- The curated catalog is also available as `OpenData_rows.csv`
- **Why choose this?** If you have a compelling research question that requires different data
- **Requires approval:** You must submit a formal proposal by Week 4 (Friday) for instructor approval
- See `Dataset-Options-Guide.md` for detailed requirements and proposal format

Recommendation: Unless you have a specific research interest that requires alternative data, we strongly recommend using the default REIT Master dataset. This ensures you can focus on analysis rather than data acquisition challenges.

Four Milestones

Milestone 1: Data Pipeline (50 points, Due Week 5 - Feb 20)

Deliverable: Python script that fetches, cleans, and merges data into analysis-ready format

Key Requirements:

- Load REIT Master dataset (All REITs, 120+ months) - **Required**
- Fetch and integrate supplementary data (e.g., FRED economic indicators, housing data, labor market data, or other datasets from Orbis Open Dataset Catalog) - **Optional but recommended**
- Clean missing values with documented decisions
- Merge datasets on Date/Month
- Output tidy panel structure (Entity=REIT, Time=Month)
- Save as CSV with metadata documentation

Skills Practiced: pandas, data wrangling, API integration, reproducible pipelines

---

Milestone 2: EDA Dashboard (50 points, Due Week 9 - Mar 27)

Deliverable: Jupyter notebook with comprehensive exploratory data analysis

Key Requirements:

- Correlation heatmap: interest rates vs. REIT returns
- Lagged effect analysis: how long do rate changes affect returns?
- Sector segmentation: identify "sensitive" vs. "resilient" REIT sectors
- Time series decomposition: trend, seasonality, residuals
- Summary statistics by REIT and time period
- Formulate testable hypotheses for econometric models

Skills Practiced: Visualization, EDA, hypothesis generation, economic intuition

---

Milestone 3: Econometric Models (50 points, Due Week 12 - Apr 17)

Deliverable: Python script with panel regression models, diagnostics, and robustness checks

Key Requirements:

- **Model A:** Fixed Effects regression controlling for REIT characteristics
  - Specification: `Return ~ Economic_Indicators + REIT_Characteristics + REIT_FE + Time_FE`
  - Clustered standard errors
- **Model B:** One additional specification (choose one):
  - Difference-in-Differences (policy shock analysis)
  - ARIMA time series forecast
  - Machine Learning comparison (Random Forest vs. OLS)
- Full diagnostics: heteroskedasticity tests, multicollinearity (VIF), outlier treatment
- Robustness checks: alternative lag structures, placebo tests, sensitivity analysis
- Publication-ready regression tables

Skills Practiced: Panel data, causal inference, model diagnostics, specification testing

---

Milestone 4: Final Investment Memo (50 points, Due Week 14 - May 1)

Deliverable: Professional memo (5-7 pages) + individual addendum (1 page)

Team Memo Components:

1. **Executive Summary** (0.5 page): Key finding + investment implication
2. **Methodology** (1 page): Data sources, panel structure, model equations
3. **Results** (1.5 pages):
   - Table 1: Fixed Effects model
   - Table 2: Alternative specification
   - Figure 1: Key visualization (e.g., sector returns over time)
   - Figure 2: Diagnostic plot (residuals vs. fitted)
   - Economic interpretation of coefficients
4. **Conclusions & Recommendations** (1 page): Buy/Hold/Sell recommendations with caveats
5. **References & AI Audit Appendix** (0.5-1 page)

Individual Addendum (1 page per student):

- Personal contribution (2-4 bullets)

- One methodological decision you would defend (2-4 sentences)
- One key limitation (2-4 sentences)
- AI audit notes for your work

Skills Practiced: Professional communication, technical translation, investment reasoning

---

Final Presentation (100 points, Weeks 14–15)

Format: "Investment Committee" Simulation **Duration:** 8 minutes pitch + 2 minutes Q&A per team
Audience: Instructor, peers, optional guest expert (REIT analyst or portfolio manager)

Structure:

1. **Problem Statement** (1 min): Why this analysis matters
2. **Data & Methods** (2 min): REIT panel + econometric approach
3. **Key Results** (4 min): Main finding with visual evidence
4. **Implications** (2 min): Investment recommendations
5. **Caveats** (1 min): Assumptions and limitations

Q&A Examples:

- "Why Fixed Effects over Pooled OLS?"
- "How confident are you in this estimate given $R^2 = 0.35$?"
- "What if parallel trends are violated?"

Grading Breakdown

*Percentages are approximate; **points** are the authoritative grading basis.*

Item	Points	Weight (of 1,450)	Due Date
M1: Data Pipeline	50	3.5%	Week 5
M2: EDA Dashboard	50	3.5%	Week 9
M3: Econometric Models	50	3.5%	Week 12
M4: Final Memo	50	3.5%	Week 14
Final Presentation	**100**	**6%**	Weeks 14–15
Total (Milestones + Presentation)	**300**	**~21%**	

Dataset: REIT Master Dataset (Absolute Default)

All students use the instructor-provided REIT dataset for Weeks 1-13. This ensures:

- Consistent grading across all teams
- Peer support and collaboration
- Fast feedback and auto-grading alignment

Default Track:

1. **REIT Master dataset** (Required): Monthly returns and characteristics for 500+ US equity REITs (monthly, REIT-level)
    - Variables: `permno`, `ym`, `ret`, `mcap`, `sector`, `price`
    - Source: CRSP/Ziman Real Estate Database
2. **Supplementary Data** (Optional - Your Choice): You can enhance your analysis by adding other datasets. Common examples include:
    - **FRED Economic Data**: Interest rates, inflation, unemployment (fetch via `pandas-datareader` API; series like `FEDFUNDS`, `MORTGAGE30US`, `CPIAUCSL`, `UNRATE`)
    - **Other datasets from Orbis Open Dataset Catalog**: Housing data, labor market indicators, climate data, etc.
    - **Other public data sources**: Any publicly available data that complements your research question

Capstone Alternative Datasets (By Exception Only):

- Teams may propose an alternative dataset from the **Orbis Open Dataset Catalog** at phdai.ai (also available as `OpenData_rows.csv`)
- Requires formal proposal and instructor approval by Week 4
- Proposals granted sparingly; default to REIT track unless compelling research reason
- See `Dataset-Options-Guide.md` for requirements

Team Formation Guidelines

Team Size: 3-4 students **Formation Deadline:** Monday, Week 4 (February 9, 2026)

How to Form Teams:

1. **Self-Selection (Preferred):** Find teammates with complementary skills
    - Consider: coding experience, econometrics background, writing ability, time availability
2. **Instructor Assignment (if needed):** Students not in a team by Monday, Week 4 (February 9) will be assigned
3. **Team Diversity:** Mix of skills is ideal (e.g., one strong coder + one strong writer + one econometrician)

Team Dynamics Best Practices:

- Establish weekly meeting time (outside class)
- Use version control (GitHub) for code collaboration
- Assign roles: Data Engineer, Analyst, Writer, Reviewer (rotate as needed)
- Document individual contributions throughout semester (required for M4)

Conflict Resolution:

- If team issues arise, contact instructor immediately
- Peer evaluation form (due end of Week 15) weights individual contributions
- Severe team dysfunction: instructor may adjust individual grades

---

Timeline and Checkpoints

Week	Capstone Activity	Deliverable
1-3	Team formation, research question refinement	Team roster (due Mon, Week 4)
3-4	Data exploration, pipeline design	Draft cleaning script
5	**Milestone 1 Due**	Data Pipeline (50 pts)
6-7	EDA, correlation analysis, visualization	Draft notebook
8-9	Hypothesis formulation, sector analysis	EDA insights
9	**Milestone 2 Due**	EDA Dashboard (50 pts)
10-11	Model specification, FE/DiD estimation	Regression outputs
12	**Milestone 3 Due**	Econometric Models (50 pts)
13	Diagnostics, robustness checks, alternative specs	Model comparison
14	**Milestone 4 Due + Presentations Day 1**	Final Memo (50 pts) + Presentation (Day 1)
15	**Presentations Day 2 + Peer Evaluations Due**	Presentation (Day 2) + Peer evaluation

Feedback Loop:

- Instructor provides written feedback within 1 week of each milestone
- Office hours available throughout for debugging and methodology questions
- Peer review sessions in Weeks 6, 10, 14

Success Criteria

What does "A" work look like?

Technical Excellence

- Code runs end-to-end without errors (reproducibility)
- Models are economically sensible and statistically rigorous
- Diagnostics address all Gauss-Markov assumptions
- Robustness checks strengthen causal claims

Analytical Depth

- Clear economic intuition guides specification choices
- Interpretation connects coefficients to real-world mechanisms
- Limitations and caveats are thoughtfully discussed
- Results answer the original research question

Communication Quality

- Executive summary is clear and compelling (non-technical audience)

- Tables and figures are publication-ready (titles, labels, legends)
- Methodology is transparent and replicable
- Recommendations are evidence-based with honest uncertainty

Professional Conduct

- All deadlines met
- Team collaboration is effective and equitable
- AI Audit Appendix demonstrates responsible AI use
- Individual addendum shows genuine reflection and accountability

---

Common Pitfalls to Avoid

Data Pipeline (M1)

- ❌ Hardcoded file paths (`C:\Users\...`)
- ❌ No documentation of missing value decisions
- ❌ Merge produces duplicate rows or data loss
- ☑ Use relative paths, document all cleaning choices, verify row counts

EDA Dashboard (M2)

- ❌ Unlabeled plots, no titles or axis labels
- ❌ No economic interpretation, just mechanical descriptions
- ❌ Ignoring outliers or data quality issues
- ☑ Publication-ready visuals, connect patterns to economic theory

Econometric Models (M3)

- ❌ Blindly accepting AI-generated specifications
- ❌ Skipping diagnostics or ignoring violations
- ❌ Interpreting correlation as causation
- ☑ Test assumptions, apply robust fixes, defend causal identification

Final Memo (M4)

- ❌ Overly technical jargon for non-economist readers
- ❌ No investment implications or actionable recommendations
- ❌ Missing AI Audit Appendix or shallow reflection
- ☑ Translate statistics to business language, provide clear recommendations

---

Resources and Support

Office Hours

- **Instructor:** Dr. Seagraves, Monday & Wednesday 3:00-5:00 PM (Helm 122-D)
- **TAs:** See Blackboard announcements for current TA assignments and office hours.

Technical Help

- **Course GitHub Discussions:** Post questions, share debugging tips
- **CSAS Tutoring:** Free academic support, statistical consulting
- **Python Documentation:** pandas, statsmodels, linearmodels

Example Materials

- Sample milestone submissions (from previous semesters, anonymized)
- Code templates in each milestone's `starter/` folder
- Rubrics for transparent grading expectations

Academic Integrity and AI Use

AI Tools are Encouraged, but you must follow the **"Disclose, Verify, Critique"** framework:

1. Disclose

- Document all AI use in the **AI Audit Appendix** (required for all milestones)
- Specify: Which tool? What prompt? What output?

2. Verify

- Test all AI-generated code on your data
- Cross-check econometric interpretations against course material
- Run diagnostics to validate AI suggestions

3. Critique

- Identify when AI is wrong (common for econometrics)
- Explain how you corrected errors
- Demonstrate understanding beyond copy-paste

Example AI Audit Entry:

> *"Used ChatGPT to debug a pandas merge error. Prompt: 'KeyError on ticker when merging REIT data.' AI suggested using `left_on='permno', right_on='ticker'`. I verified this fixed the issue by checking row counts before/after merge. However, AI initially suggested `how='inner'` which would have dropped unmatched rows; I changed to `how='left'` to preserve all REIT observations."*

No AI Audit Appendix = Zero credit for the assignment.

FAQ

Q: Can we use a different dataset?

A: Only for the capstone, and only with instructor approval by Week 4. Weekly assignments (Weeks 1-13) must use the REIT default dataset. See `Dataset-Options-Guide.md` for proposal requirements.

Q: What if a team member doesn't contribute?

A: Document this in your peer evaluation form (due end of Week 15). Instructor may adjust individual grades if contribution disparity is severe. Contact Dr. Seagraves immediately if issues arise.

Q: How much Python experience do we need?

A: None! The course teaches Python from scratch. Weekly labs build skills progressively. By M1 (Week 5), you'll be ready for the data pipeline.

Q: Can we use Stata or R instead of Python?

A: No. Python is the required language for this course. All auto-grading, starter code, and support materials are Python-based.

Q: What if we can't meet a milestone deadline?

A: Contact Dr. Seagraves at least 48 hours in advance. Late penalty: 10% per day, up to 3 days. After 3 days, no credit unless excused absence.

Q: How are presentation grades calculated?

A: Each team receives a team score (out of 100 points) based on the rubric in `presentation_guidelines.md`. Individual adjustments may apply based on peer evaluation.

Q: What is the expected time commitment?

A: Approximately 50-100 hours over the semester (average 3-6 hours/week). Peaks during milestone weeks (Weeks 5, 9, 12, and 14–15).

Next Steps

1. **Weeks 1-3:** Form your team (use `team_formation.md` for guidance) and **choose your dataset** (default REIT Master or propose alternative from Orbis Open Dataset Catalog). **Team roster due Monday, Week 4 (February 9).**
2. **Week 2-3:** Review M1 specification in `M1-Data-Pipeline/README.md`
3. **Week 3-4:** Start exploring your chosen dataset, outline data cleaning strategy
4. **Week 4:** Draft data pipeline script, attend office hours for feedback. **If proposing alternative dataset, submit proposal by Friday.**
5. **Week 5:** Submit M1 (Data Pipeline) by Friday 11:59 PM

Good luck! This capstone is challenging, but you'll emerge with portfolio-ready work and real analytical skills.

Document prepared for QM 2023: Statistics II, Spring 2026, University of Tulsa Contact: Dr. Cayman Seagraves (cayman-seagraves@utulsa.edu)