Milestone 1: Data Pipeline - Blackboard Assignment Description

Course: QM 2023: Statistics II / Data Analytics
Due: Wednesday, February 25, 2026 by 11:59 PM
Points: 50 (25% of capstone grade)
Submission: Team submission (one per team)

Overview

Build your capstone project's foundational data infrastructure by integrating your primary dataset (REIT returns by default, or your approved alternative) with supplementary economic/market data into a clean, analysis-ready panel dataset.

Real-world context: Professional analysts spend 60-80% of their time on data engineering. A robust pipeline ensures all subsequent analysis runs smoothly and reproducibly.

What You're Submitting

1. GitHub Repository Link

Required: Submit the URL to your team's GitHub repository (created via GitHub Classroom).

Example: `https://github.com/Dr-Seagraves/qm2023-capstone-team-name`

Your repository should contain all M1 deliverables (see below).

Repository Structure

Your capstone repository should follow this modular structure:

```
QM-2023-Capstone-Repo/
├── code/
│   ├── config_paths.py              # Path management (provided)
│   ├── fetch_[dataset1]_data.py     # Fetch + clean primary dataset
│   ├── fetch_[dataset2]_data.py     # Fetch + clean supplementary dataset
│   ├── fetch_[dataset3]_data.py     # (Optional) Additional dataset
│   └── merge_final_panel.py         # Merge processed → final
├── data/
│   ├── raw/                         # Original downloaded data
│   │   ├── *_raw.csv
│   │   └── (additional raw files)
│   ├── processed/                   # Cleaned individual datasets
│   │   ├── *_clean.csv
│   │   ├── *_clean.csv
│   │   └── *_clean.csv
```

```
│   └── final/                        # Analysis-ready merged panel
│       ├── [dataset]_analysis_panel.csv  # Final merged dataset
│       └── data_dictionary.md            # Variable definitions
├── results/
│   ├── figures/
│   ├── reports/
│   └── tables/
├── tests/
│   └── .gitkeep
├── README.md                          # Team info, research question
├── M1_data_quality_report.md         # Data quality documentation
└── AI_AUDIT_APPENDIX.md              # AI disclosure (REQUIRED)
```

Data Pipeline Flow:

1. **Fetch scripts** (`fetch_*.py`) → Download/load data → Clean → Save to `data/processed/`
2. **Merge script** (`merge_final_panel.py`) → Combine processed datasets → Save to `data/final/`

Required Deliverables (in your GitHub repo)

1. Python Scripts (Modular Pipeline)

Multiple fetch scripts — One per dataset:

- `code/fetch_reit_data.py` (or your primary dataset)
- `code/fetch_fred_data.py` (or your supplementary data)
- Additional `fetch_*.py` scripts as needed

One merge script:

- `code/merge_final_panel.py` — Combines all processed datasets into final panel

Requirements:

- Must run without errors using relative paths only (via `config_paths.py`)
- Clear section headers and comments
- Prints before/after row counts and summary statistics

---

2. Root README.md (Project Overview)

Location: Repository root

Must include:

- Team members and roles
- Research question (1-2 sentences)
- Dataset overview (primary + supplementary sources)
- Preliminary hypotheses (3+)
- How to run the pipeline (step-by-step)

---

3. Final Dataset: `data/final/[dataset_name]_analysis_panel.csv`

- Clean panel dataset in **long format** (Entity × Time)
- Shows outcome variable(s), entity characteristics, and 10-15+ supplementary variables
- No missing keys; properly merged; ready for regression analysis

---

4. Data Dictionary: `data/final/data_dictionary.md`

- Dataset overview (N entities, N time periods, date range)
- Variable definitions table (variable, description, type, source, units)
- Cleaning decisions summary

---

5. Data Quality Report: `M1_data_quality_report.md`

Comprehensive documentation of:

- Data sources (primary + supplementary)
- Cleaning decisions with before/after counts and economic justification
- Merge strategy and verification
- Final dataset summary with sample statistics
- Reproducibility checklist
- Ethical considerations (what data are we losing?)

---

6. AI Audit Appendix: `AI_AUDIT_APPENDIX.md`

- Documentation of all AI tool use following "Disclose-Verify-Critique" framework
- **Required:** Missing AI Audit = automatic 0/50 points

What You Should Have Ready

By this milestone, your team should have established:

☑ **Your Dataset:** Primary data source identified and loaded (REIT Master for default track, or approved alternative)

☑ **Supplementary Data:** 10-15+ economic indicators, policy measures, or market factors integrated (FRED API, custom sources, etc.)

☑ **Research Direction:** Preliminary research questions that will guide your M2 exploratory analysis and M3 econometric models

☑ **Clean Data Pipeline:** Reproducible script that handles missing values, outliers, duplicates, and merges data correctly

☑ **Panel Structure:** Long-format dataset (one row per entity-time observation) ready for panel regression

---

How to Submit

1. **Push all files to your GitHub repository** (created via GitHub Classroom)

```
git add code/*.py
git add data/raw/*.csv data/processed/*.csv data/final/*.csv data/final/*.md
git add README.md M1_data_quality_report.md AI_AUDIT_APPENDIX.md
git commit -m "M1 submission: Data Pipeline complete"
git push origin main
```

2. **Submit to Blackboard:**
   - Paste your GitHub repository URL in the text box
   - Verify all files are visible on GitHub.com before submitting
   - One submission per team (team member names should be in all documents)

Grading Criteria (50 points total)

Component	Points	Key Criteria
Reproducibility	15	Scripts run without errors; modular structure; relative paths only; clear comments
Data Cleaning	12	Missing values, duplicates, outliers handled; before/after counts documented
Merge Integrity	8	No row loss/duplication; supplementary data aligned correctly
Panel Structure	8	Correct entity-time long format; dimensions verified
Documentation	7	README, data dictionary, and data quality report complete; cleaning decisions justified

Zero-Credit Conditions:

- Missing `AI_AUDIT_APPENDIX.md` = **0/50 points**
- Script won't run (syntax/path errors) = **0/50 points**
- Hardcoded absolute paths = **-10 points**

Late Policy

- **10% penalty per day** (24-hour periods) for up to 3 days
- After 3 days: no credit accepted without prior arrangement

Resources

- **Full M1 Specification:** See `M1-Data-Pipeline/README.md` in the Capstone Project folder on Blackboard

- **GitHub Workflow Guide:** See `GitHub-Workflow-Guide.md` for team collaboration best practices
- **Starter Code:** Available in your GitHub Classroom repository
- **Office Hours:** Dr. Seagraves, Mon & Wed 3-5 PM (Helm 122-D)

Questions?

Contact Dr. Seagraves via email (cayman-seagraves@utulsa.edu) or come to office hours. Don't wait until the last day to ask questions about data sources, merge issues, or reproducibility requirements.

Remember: This pipeline is the foundation for your entire capstone project. Invest the time now to get it right, and M2/M3/M4 will go much smoother.