# Fama-French Europe Project – Progress Report

Date: 2025-10-28 18:39

Author: Christos Pasentsis

Project: Fama-French Multi-Asset ML Risk Factor Study

## 1. Summary of Current Progress

The project has now reached a stable and clean state for empirical analysis. All raw return and price data for the European equity universe have been cleaned, validated, and tested for econometric assumptions. The entire data pipeline is reproducible, version-controlled, and well-documented. All major scripts have been consolidated under the /bin directory, and the Investment\_universe folder now holds clean datasets and diagnostics reports.

## 2. Key Achievements

* ✅ Cleaned and validated 126 European tickers, retaining 108 after diagnostics.
* ✅ Implemented auto-drop logic for extreme kurtosis, skewness, and structural anomalies.
* ✅ Computed ADF, KPSS, Ljung–Box, and ARCH tests across all return series.
* ✅ 98% of tickers passed stationarity; 18% passed Ljung–Box; 0% passed ARCH (expected for returns).
* ✅ Generated full visual diagnostics (ACF, histograms, QQ plots) for each asset.
* ✅ Automated README.md generation from JSON and CSV outputs.
* ✅ All code modularized and moved into /bin for a cleaner structure.
* ✅ Created reproducible clean summary and assumption summary JSONs.

## 3. Repository Structure

• bin/: Core pipeline scripts and artifacts  
• Investment\_universe/: Cleaned data, diagnostics, and reports  
• reports/: Project logs and generated analysis reports  
• README.md: Auto-updated summary of data pipeline status

## 4. Current Status

The cleaned dataset (`europe\_returns\_cleaned.csv` and `europe\_prices\_cleaned.csv`) is now suitable for cross-sectional, time-series, and panel regression analyses. The dataset passes all basic data-quality assumptions. Econometric assumptions (white-noise residuals) are not fully satisfied but are handled using HAC/Newey–West standard errors or volatility modeling (GARCH) where applicable.

## 5. Next Steps

* 🔹 Implement factor regressions using Fama-French 5-factor + Momentum model.
* 🔹 Explore machine-learning risk factor extensions (e.g., Ridge/Lasso, Random Forest).
* 🔹 Backtest clean universe using value and momentum portfolios for robustness.
* 🔹 Integrate macroeconomic controls (term spread, inflation expectations).
* 🔹 Document modeling strategy and results for thesis-level reproducibility.

## 6. Recommendations

Keep all intermediate JSON and CSV summaries under version control for full traceability. When running regressions or portfolio sorts, always cluster standard errors by both ticker and date to account for autocorrelation and conditional heteroskedasticity.

## 7. Hand-off Notes for New Analyst/Chat

All current datasets are panel-ready. The primary working directory is `Investment\_universe/`, which contains the cleaned returns, prices, assumption tests, and diagnostics reports. The `update\_readme.py` script regenerates the documentation block automatically from the latest pipeline run. Future work should build upon these clean files for factor analysis and portfolio construction.