YFinance Core Tables Pipeline

This notebook extracts and processes real financial data using the Yahoo Finance API (`yfinance`) and organizes it into structured DataFrames for ESG analysis and fund fact sheet generation. It produces multiple interrelated tables, covering security metadata, ESG scores, historical prices, benchmark data, and synthetic portfolio construction. All tables are prepared for Snowflake ingestion, supporting scalable, enterprise-ready ESG reporting.

Execution Instructions

1. Update your Snowflake credentials in the `.env` file or secure config.
2. Run the notebook sequentially from top to bottom to avoid dependency errors.
3. Use `load_to_snowflake_merge()` for table ingestion to avoid duplication.

File Roadmap

- **Scrape major US index constituents** (S&P 500, NASDAQ 100, Dow Jones)
- **Query Yahoo Finance** for ESG scores and 10 years of historical prices
- **Construct performance and ESG data table** (`df_performance`) with `data_type` flag
- **Generate Security Master** metadata (`df_security_master`)
- **Create synthetic ESG fund holdings** and simulated performance
- **Build benchmark reference tables** using index ETFs (e.g., SHE, ENRG)
- **Validate and inspect all final DataFrames**
- **Prepare all outputs for Snowflake ingestion**

Output Summary

DataFrame Table Name	Description	Source	Snowflake Table Name	Transfer to Snowflake Status
DF_SECURITY_MASTER	Company metadata: sector, industry, market cap	yFinance	SECURITY_MASTER	Completed
DF_PERFORMANCE	ESG and price data tagged by `data_type`	yFinance	SECURITY_PERFORMANCE_HISTORY	Completed
DF_HOLDINGSDETAILS	Simulated ESG fund holdings: tickers, weights, market value	Synthetic	HOLDINGSDETAILS	Not Yet
DF_PORTFOLIOPERFORMANCE	Portfolio-level ESG scores and metadata	Synthetic	PORTFOLIOPERFORMANCE	Not Yet
DF_BENCHMARK_PERFORMANCE	Historical ETF/index prices for benchmarking	yFinance	BENCHMARKPERFORMANCE	Not Yet
DF_BENCHMARK_GENERALINFO	Metadata about benchmark indices	yFinance	BENCHMARKGENERALINFORMATION	Not Yet
DF_BENCHMARK_CHARACTERISTICS	Valuation & fundamental metrics (e.g., PE, market cap)	yFinance	BENCHMARKCHARACTERISTICS	Not Yet

Next Steps & Future Improvements

Security Master + ESG Performance Table

- Add intraday pricing support to better align with real-world execution times
- Introduce dynamic metadata patching if a ticker appears mid-pipeline
- Implement fallback or imputation for missing ESG scores using:
    - Country-level or sector-level averages
- Extend performance history or cache for more robust backfilling
- Improve error handling and retry logic to reduce missing records

Synthetic ESG Fund Construction

- Automate creation of portfolios using blended ESG metrics or dynamic clustering
- Introduce ESG attribution to show score drivers at the holding level
- Add time-series ESG tracking to monitor portfolio ESG trend changes
- Load `HOLDINGSDETAILS` and `PORTFOLIOPERFORMANCE` into Snowflake with merge keys
- Evaluate integration of actual fund holdings (if available) for real-world scoring

Benchmark Integration

- Create a Snowflake ingestion pipeline for all benchmark DataFrames
- Extend benchmark return metrics to include rolling YTD, 1Y, and 3Y returns
- Create a formal benchmark-to-portfolio mapping table for joins
- Add non-ESG benchmark ETFs (e.g., SPY, ACWI) for relative performance context
- Centralize benchmark definitions to allow scalable mapping across new funds

Each next step aligns with **pipeline scalability**, **fact sheet completeness**, and **production-readiness goals**. Focus areas include:

- More realistic performance simulation
- Robust ESG coverage
- Better join logic for benchmark comparison
- Cleaner, repeatable Snowflake integration