Task 2

Project title: Private Market Conditions Index

Team member: Richard Mihhels

Identifying business goals

Background

Financial markets are strongly influenced by macroeconomic conditions, but in practice it is often unclear which metrics matter most for different types of assets. Private equity and private credit are especially sensitive to growth, interest rates, credit spreads and risk sentiment, yet most analysis is based on narratives rather than systemic evidence. For this project, I will focus on US markets (with the option to extend, when I have time, to other markets like Europe and Asia), using public market proxies for private equity and private credit, together with macro and financial time series from FRED and other open data sources.

Business goals

From a business point of view, the main stakeholder is a research-oriented investment analyst (at a fund, family office or macro research team) who wants to understand the context going forward.

Which macroeconomic and financial variables are most correlated with the performance of private equity proxies (listed Private Equity firms, small-cap value ETFs) and private credit proxies (high yield bond ETFs, loan ETFs, BDC ETFs). Whether these relationships are contemporaneous or show lead lag structure (do changes in credit spreads or policy rates lead returns?). How stable these correlations are over time and across different regimes (low vs high rates, stressed vs calm credit conditions).

This project will not build a trading system, it focuses on analytics and understanding, not signal production or execution.

Business success criteria

The project is considered a success when these criteria are accomplished.

- It produces a clear, documented correlation map between key macro/credit variables and the return of the chosen asset proxies.
- It generates interpretable visualizations (heatmaps, rolling correlations, scatter plots) that a non-technical finance person could use in a slide deck.
- It identifies a short list of key drivers (high yield spreads, policy rate changes, unemployment, inflation) with consistent and economically meaningful relationships to asset returns.
- All code and documentation are reproducible from the repository using public data.

No direct business revenue is expected. The value is in improved analytical insight and future reusability of the codebase.

Assessing the situation

Inventory of resources

- Public data: Yahoo Finance, FRED economic data.
- Tools: GitHub repository, Python (pandas, numpy, matplotlib, scipy), Jupyter notebooks.
- Computing: personal laptop, no special hardware needed.

Requirements, assumptions and constraints

- All datasets must be freely available and legally usable for academic purposes.
- The project will focus primarily on US data to keep scope manageable.
- I assume FRED/Yahoo data is sufficiently accurate for research and that any missing data can be handled via standard methods.

Risks and contingencies

- Some series may be discontinued or have short histories, contingency: replace with alternative series or reduce the set of indicators.
- Data quality issues (missing values, structural breaks) may reduce sample size, contingency: document these and run sensitivity check.
- Correlations may be weak or unstable, success does not depend on strong results, only on rigorous. Methodology and clear documentation.

Terminology

- Private equity proxy: listed firms and ETFs whose performance broadly reflects private equity (PSP, BX and so on).

- Private credit proxy: HY bond ETFs, loan ETFs, BDC ETFs (HYG, BKLN, BIZD).
- Macro variables: growth, labour, inflation, rates, credit spreads, financial stress, volatility.
- Market movement: monthly total returns of the chosen asset proxies.

Costs and benefits

- Costs: student time.
- Benefits: reusable code to quickly explore macro-market relationships, nicer understanding of which environment tends to help or hurt specific asset classes and a structured template for future research projects.

Defining data mining goals

Data mining goals

- Build a clean, aligned monthly panel of economic indicators and asset returns for the US.
- Quantify relationships using static correlation, rolling correlation over time, simple linear models to check robustness.
- Identify the most relevant variables for each group and describe the sign and rough strength of the relationships.

Data mining success criteria

- At least 5-10 macro/credit variables with properly documented transformations (levels, changes, YoY, standardisation).
- At least 8-12 asset proxies (mix of private equity, private credit and benchmarks) with monthly return series.
- Correlation tables and plots that are reproducible from the repo, easy to interpret(consistent naming and legends), accompanied by short written interpretations.

Task 3

Gathering data

Outline data requirements

To study correlation between economic conditions and market movements, I need:

1. Asset returns:
    - Daily prices for a basket of US-listed ETFs and stock representing: private equity (PSP, BX, KKR, APO, CG, VBR), private credit (HYG, JNK, BKLN, SRLN, BIZD, ARCC), benchmarks (SPY, IEF).
    - These will be transformed into monthly total returns.
2. Economic and financial indicators:
    - Growth and activity: GDP growth, industrial production, payrolls, real retail sales.
    - Labour market: unemployment rate, participation rate.
    - Inflation: CPI headline and core, possibly PCE.
    - Interest rates and curve: Fed Funds rate, 2y and 10y Treasury yields.
    - Credit & stress: HY and IG spreads, financial stress index.
    - Risk sentiment: VIX.

    I require at least 15-20 years of data to cover multiple cycles.

Verify data availability

- Yahoo Finance provides historical OHLCV for all the ETFs and stocks via API,
- FRED exposes all listed series through a free API (API key needed) and web CSV downloads.
- Preliminary checks show that most chosen series have data from early 1990s or 2000s onwards.

Define selection criteria

- Only US-focused series will be used initially to avoid mixing regional effects.
- For a series to be included it should have at least 10 years of history overlapping with the asset data and it should be clearly interpretable and relevant to growth, inflation, rates, credit or risk sentiment.
- If two series are almost duplicates (multiple CPI variants), I will pick the one with the clearest interpretation.

Describing data

Once downloaded, I will produce a brief description of each dataset.

- Assets data: numbers of tickers, date range, frequency, missing days. For each asset: mean monthly return, standard deviation, min/max, count of observations.
- Macro/FRED data: For each series: name, unit, transformation used (level, log, diff, YoY), date range and frequency (monthly vs quarterly).
- Summary statistics (mean, std, min/max, number of NAs).

I will create a simple data dictionary (table) explaining each variable in plain language, "BAMLH0A0HYM2 – US high yield option-adjusted spread, higher values mean tighter credit conditions for risky borrowers".

Exploring data

Exploration will focus on:

- Time series plot: Asset price levels and monthly returns (SPY vs PSP vs HYG and so on). Macro series over time (inflation, unemployment, HY spreads, stress index).
- Distributions: Histograms / kernel density plots for monthly returns by asset group (PE vs PC vs benchmark). Histograms for key macro variables and their changes (ΔHY spreads, ΔFed Funds, CPI YoY).
- Basic relationships: Scatter plots of asset returns vs selected macro features for a few long samples periods (HYG returns vs HY spread changes, PSP returns vs GDP growth).

This step will help identify obvious data issues (outliers, structural breaks) and give intuition before formal correlation analysis.

Verifying data quality

I will check data quality along several dimensions:

- Missing values and gaps: For each series, count Nas and detect periods with no data. For macro series with quarterly frequency, verify that forward-filling to monthly does not create unrealistic jumps.
- Consistency and alignment: Ensure that all series are aligned to month-end dates before analysis. Confirm that there are no duplicate timestamps or overlapping intervals.
- Outliers: Use simple rules (values more than +-5 standard deviations) to flag extreme observations. Verify whether outliers correspond to real events (2008 crisis, COVID shock) rather than data errors.

- Unit and transformation checks: Confirm that percentage series (HY OAS, unemployment, Fed Funds) are interpreted correctly (in % points vs decimals).
- Document all transformations (log, diff, YoY, standardization) to keep the pipeline reproducible.

If serious quality problems are found for any series (very short history or structural breaks), I will either remove it from the final feature set or clearly label it as experimental.

Task 4

Task list and time plan

Team: 1 member – Richard Mihhels. Approximately 60 hours of work.

1. Repository and environment setup (3 h)
    - Create Github repo, invite instructors, set up basic folder structure and requirements file.
2. Data collection & documentation (10 h)
    - Implement scripts to download Yahoo Finance and FRED series.
    - Write a short data dictionary and save raw data to the repo.
3. Data preparation & feature engineering (15 h)
    - Convert daily prices to monthly returns.
    - Align macro data to monthly frequency and create transformed features (YoY, changes, z-scores).
4. Analysis & visualization (20 h)
    - Compute static and rolling correlations, explore simple regression.
    - Generate heatmaps, rolling correlation plots and scatterplots and write short interpretations.
5. Reporting & polishing (12 h)
    - Prepare the CRISP-DM report sections, final figures and repository README.
    - Ensure code is reproducible and clearly commented.

Methods and tools

- Tools: Python, Jupyter, pandas/numpy, matplotlib/seaborn, statsmodels/scikit-learn, Github.
- Methods: time-series resampling, correlation analysis (Pearson/Spearman), basic linear regression, visual exploration (heatmaps, rolling windows).

Comments: focus is on interpretability and robustness, not complex machine learning. The main deliverable is a reusable analytical framework and a clear empirical summary of macro-market relationship.