Make Data Model Ready

1. Past and current week's work

Before this week we built a clean foundation. We sourced firm-level adjusted price and dividends, metro-level CPI and unemployment, and the ten-year Treasury yield. We normalized names and units, converted percentages to decimal rates, and harmonized dates to a monthly cadence. We merged macro series to firms by month and metro with a left join anchored on the firm timeline so the financial target stays aligned. We engineered month-over-month and year-over-year changes for the macro series because both level and momentum can matter in downstream models. Exact duplicates were removed earlier, and we created small helpers for safe reading and writing. That scaffolding lets Week 4 focus on disciplined preprocessing rather than plumbing.

The goal this week was to transform the merged tables into a modeling matrix that is numerically stable, time-safe, and easy to reproduce. The steps below strictly follow our notebook under the Week 4 header, with the same sequence and scope.

1) Missing values

We began with a complete audit of nulls for the core numeric series. For CPI, its year-over-year rate, month-over-month change, unemployment rate, and unemployment month-over-month change, we filled within each metro using forward fill after sorting by metro and date. This respects the structure of monthly economic series, where the latest observation in a metro is a reasonable stand-in for the next if one value is missing. To avoid leaving holes at the start of a metro history or after rare multi-month gaps, we then replaced any residual missing entries with the median of that series computed over the available data. We did not create synthetic labels at any point. This two-step rule balances continuity with conservatism: the metro-level forward fill keeps local trajectories intact, and the median catch-all prevents stray NaNs from leaking into later steps while avoiding aggressive interpolation.
We opted for metro-level forward fill because many official metro series are published with occasional lags or single-month holes. Forward fill preserves the persistence that monthly CPI and unemployment typically show, especially in levels. The median fallback is deliberately simple and robust. It prevents the creation of extreme values that a parametric imputation could introduce, and it touches only the handful of entries that forward fill cannot address. This approach is easy to explain and repeat, which matters for transparency.

2) Outlier handling

We controlled extremes with a mix of domain limits and a robust rule on change features. Unemployment was clipped inside a wide logical band from zero to twenty-five percent. The ten-year yield was clipped inside zero to fifteen percent. Trailing-twelve-month dividends were clipped at the ninety-ninth percentile to curb the influence of extraordinary distributions. For high-volatility change features we applied an interquartile-range based winsorization. Specifically, we computed the first and third quartiles and clipped values outside Q1 minus one and a half times IQR and Q3 plus one and a half times IQR. We applied this rule to CPI month-over-month, CPI year-over-year, unemployment month-over-month, and the month-over-month change in the ten-year yield.

The domain clips are wide enough that they spare normal months but prevent numerical explosions triggered by rare events or reporting glitches. Using an IQR threshold on changes tempers only the most extreme tails while preserving the center and ordering of the series. Change features carry signal during shocks, so a light touch is better than filtering. After this step, distribution summaries in the notebook show medians unchanged, hinges close to their original values, and only the far tails compressed. That is the pattern we want.

3) Transformations and standardization

We reduced skew in financial variables by constructing adj_price_log and dividend_ttm_log with a log-plus-one transform. These become the financial versions used for modeling. We then standardized continuous variables to better condition linear and distance-based learners. In an early diagnostic, we fitted a StandardScaler on the full dataset to inspect how means and standard deviations shift after scaling and saved that scaler for reference. Our final preprocessing for the modeling splits uses a pipeline that fits the scaler on the training window only and applies the learned parameters to validation and test. This ensures time safety. The scaled set includes CPI level, CPI year-over-year, CPI month-over-month, unemployment level, unemployment month-over-month, the ten-year yield, and its month-over-month change.

Log transforms compress multiplicative shock into an additive range that plays well with linear baselines without discarding rank information. Standardization helps convergence for methods sensitive to feature scale and makes coefficients more interpretable. The initial global scaling was a diagnostic to confirm magnitude and dispersion changes; the time-safe pipeline is what we use for train, validation, and test so that no information from the future can seep into training.

4) Feature pruning

Once log variables were created, we removed raw adj_price, dividend, and dividend_ttm from the modeling frame. We kept their log counterparts. This prevents duplicated signal and reduces collinearity risk. We left other engineered change features in place so models can learn from both stance and recent momentum.
Dropping superseded columns makes the design matrix lean and avoids unintended weighting where a raw and a transformed version of the same signal would be available. Keeping level and change features together allows the learning algorithm to pick the combinations that matter without hard-coding a choice.

5) Categorical encoding

We used one-hot encoding for the compact categorical fields. In one step we exported an encoded table using a direct get_dummies call so that we could inspect the resulting columns and feature names. For the final modeling flow we embedded OneHotEncoder in a ColumnTransformer so encoding lives inside the same time-safe pipeline as scaling. Unknown categories are ignored at transform time, and the output is dense. Encoded fields are company identifier and metro. We did not explode the date into calendar dummies in preprocessing, because our evaluation is chronological and naive month indicators can blur seasonality across splits.
The dataset's categorical cardinalities are modest, so one-hot encoding is both interpretable and safe. Including encoding inside the pipeline guarantees that the mapping from categories to columns is learned on the training window and then held fixed, which is the correct behavior for a temporal split.

6) Bucketization

We demonstrated bucketization by binning the ten-year yield into three quantile bins and CPI year-over-year into four quantile bins, then one-hot encoding those bins and exporting that exploratory table. This step is not included in the final train, validation, and test matrices.
The goal was to document how we would categorize continuous variables if a model or analysis called for it. For this assignment we keep the modeling frame continuous and reserve bins for later experiments that might benefit from coarse grouping.

7) Chronological split and final exports

We sorted the dataset by date and built a strict seventy, fifteen, fifteen split for train, validation, and test. We printed the shapes and date spans for each slice as a sanity check. We then created a preprocessing pipeline with a ColumnTransformer that applies the scaler to numeric columns and the one-hot encoder to categorical columns. The pipeline was fitted on the training slice and applied to validation and test. We

converted the results back to DataFrames with readable feature names, ran quick checks for missing values and duplicate rows, and saved three CSVs for train, validation, and test. These files are the modeling-ready artifacts for next week.
Economic variables are autocorrelated and regimes shift. Random splits can leak future information and inflate scores. A date-ordered split enforces realism. Fitting preprocessing only on the training window maintains that realism end to end.

2. Observations from the Pipeline

Running the Week 4 pipeline surfaced patterns that guide modeling and evaluation. Missingness is not random: metro-level CPI has a few early-period holes (likely later onboarding), which our metro-level forward fill covers with minimal distortion; rare residual gaps are handled by the series median. Unemployment gaps appear as isolated single months and are likewise resolved by the two-step rule without warping broader swings. After filling, completeness checks confirm all modeling columns are populated.

Outlier control behaved as intended. Wide domain clips on unemployment and the ten-year yield barely touch normal months and primarily guard against rare spikes or bad readings. Dividend tails thin visibly after the 99th-percentile clip. For change-rate features (CPI MoM/YoY, unemployment MoM, yield MoM), the IQR winsorization trims only extreme edges while preserving the sequence and shape of known shock periods; medians and hinge points remain essentially intact. Practically, this should prevent a handful of violent months from dominating loss surfaces in linear learners.

Transformations and scaling had the expected effects. Log-mapping adjusted prices and trailing dividends produced more regular distributions. A diagnostic global scaling confirmed dispersion shifts; importantly, the final train/validation/test exports use a scaler fitted only on the training window via the pipeline. Training features center near zero with unit variance; validation and test retain sensible dispersion without refitting.

Encoding remains compact and useful. One-hot variables for company and metro expand the matrix modestly, with readable, stable names. Quick linear checks show these indicators reduce residual variance in training while remaining stable on validation—evidence that firm and regional fixed effects add durable structure beyond macro variables.

Chronological splitting behaves honestly. Validation metrics are lower than training, especially around macro shock windows, which is the pattern we want when leakage is prevented. This reinforces that regime change will challenge models and that our evaluation setup is appropriately strict.

Finally, the cleaned matrix links cleanly to real-economy behavior. Higher price levels with strong monthly momentum align with household pressure and firm cost squeezes that influence margins, payouts, and valuations. Small unemployment changes often mark stress diffusing into consumption and rents—important for region-exposed firms. The ten-year yield captures monetary stance and financing conditions. Company identifiers proxy strategic choices and customer mix; metro identifiers capture local demand and policy texture that macro aggregates only partially reflect. Preserving both level and momentum under time-safe preprocessing gives downstream models a fair chance to learn relationships with clear economic meaning.

3. Conclusion and next steps

This week leaves us with a disciplined, time-safe dataset and a clear modeling runway. We resolved missingness with metro-level forward fill and a conservative median fallback, controlled extremes with domain clips and IQR winsorization, compressed skew in prices and dividends with log transforms, standardized continuous features using a scaler fit on the training window, and encoded compact categories inside the same pipeline. The final exports are three chronological slices with identical schemas and datestamped artifacts, so reruns are straightforward and comparisons will be fair. In next week's work we will anchor baselines with regularized linear models and a small gradient boosting tree to establish a reference for both accuracy and stability. We will watch two sensitivities closely: rows touched by the median fallback and the winsorization threshold on change features. If either shows outsized impact on validation error, we will document it and run a limited alternative while keeping the main pipeline unchanged for comparability. Feature importance and partial dependence will be interpreted in economic terms using the preserved level and momentum signals and the firm and metro indicators. The objective is a transparent, defensible benchmark that reflects real regimes rather than artifacts of preprocessing.