Perform Exploratory Data Analysis

1. Review of the dataset partition strategy

We built a single monthly panel that ties each public real-estate company to one metro and aligns firm variables with metro CPI and unemployment and a national 10-year Treasury series on the same month-end clock. Because these are time-ordered observations, we split by time, not at random. The merged table is sorted chronologically and then cut into three consecutive blocks: 70% for training, 15% for validation, and 15% for testing. This keeps future months out of the learning and tuning stages and gives a final slice that imitates real deployment.

The split produced the following concrete windows from the notebook run: total samples 769; training set shape 538×13 covering 2014-12-31 to 2022-05-31; validation set shape 115×13 covering 2022-05-31 to 2023-12-31; test set shape 116×13 covering 2024-01-31 to 2025-08-31. The code also writes these three tables to disk for reproducibility.

We use a single global timeline across companies, rather than cutting per company, so that all firms in a given block share the same macro backdrop. That choice simplifies model interpretation because cross-firm differences are less likely to be artifacts of mismatched periods. It also avoids look-ahead: hyperparameter selection only uses the middle block, and no inspection of the final test block occurs until the very end.

The lengths of the blocks were chosen to balance two needs. The training window is long enough to include multiple macro regimes and payout cycles, which improves parameter stability. The validation period is wide enough to expose models to nontrivial regime change after 2022, useful for selecting features and regularization strength. The test period sits in the most recent months and serves strictly as the last untouched yardstick. These decisions are transparent and easy to reconstruct because the code sorts by date, slices by index boundaries, and prints the date ranges and shapes after splitting.

2. EDA analysis and insights: what we ran and the results

All EDA is performed on the training set only to avoid leakage. The goal is to check integrity, understand distributions, and map basic dependence among predictors so we can decide on preprocessing and a first feature shortlist.

2.1 Schema and basic integrity.

The merged training table has 13 columns with the following dtypes printed by the notebook: date datetime64; company object; metro object; numeric fields are adj_price, dividend, dividend_ttm, cpi, cpi_yoy, cpi_mom, unemp, unemp_mom, ten_year, ten_year_mom. A quick preview of the top rows verifies alignment across layers. For example, the first five rows for BXP in Boston begin at 2014-12-31 with adj_price 100.757, dividend 5.80, dividend_ttm 5.80, ten_year 2.207273; by 2015-01-31, adj_price is 108.696, dividend 7.75, cpi 254.556, unemp 4.7, ten_year 1.881500, and ten_year_mom −0.325773. This confirms that firm, city, and national series are aligned on month-end.

2.2 Coverage and missingness.

Training set size is 538 rows. The describe table shows complete coverage for most numeric variables with counts of 538, while metro CPI and its changes have 351 non-null observations in training. That shorter count reflects the start dates of the metro CPI series within the training window; unemployment, ten-year yield, and their monthly changes are fully populated. The code prints a “Missing value statistics” block and plots the pattern, which shows concentration of CPI gaps in the earliest months. This guides two safe options before modeling: forward-fill within metro after the series starts, or restrict any CPI-based features to months where they exist and let the model learn on a slightly smaller but cleaner set.

2.3 Time coverage by company.

A grouped summary on the training slice reports the span and count per company: BXP 2014-12-31 to 2022-05-31 with 90 rows; ELME 2015-01-31 to 2022-05-31 with 89; EQR 2014-12-31 to 2022-05-31 with 90; REXR 2014-12-31 to 2022-04-30 with 89; SLG 2014-12-31 to 2022-05-31 with 90; TRNO 2014-12-31 to 2022-05-31 with 90. This even coverage means later models can include company or metro effects without severe imbalance.

2.4 Descriptive statistics.

On the training set, the notebook’s summary shows the following means: adj_price 54.927262, dividend 0.406259, dividend_ttm 4.578889, cpi 262.11314, unemp 5.173234, ten_year 1.976554. Change variables are centered near zero as expected: cpi_mom mean 0.003204 and unemp_mom mean −0.023234. These numbers confirm that differencing has stabilized the scale of the change features around a small central band, while level variables remain on different scales. That supports a plan to standardize predictors prior to model fitting or to use algorithms insensitive to scale.

2.5 Distribution checks.

The notebook draws histograms for all numeric fields. Prices are right-skewed with a long upper tail, dividends are very spiky around zero with occasional larger values, and the twelve-month dividend totals are smoother but still positively skewed. The change variables for CPI, unemployment, and the ten-year cluster tightly around zero with modest tails. The implication is straightforward: consider log transform or winsorization for prices and payout totals and treat change features with simple standard scaling.

2.6 Cross-sectional heterogeneity.

Boxplots compare CPI and unemployment by metro through the lens of each firm's mapped city. The spreads show that some metros experienced consistently higher inflation or tighter labor markets during the training window. That suggests fixed effects or metro dummies can absorb baseline level differences so that the model focuses on within-metro changes over time.

2.7 Dependence structure.

A Pearson correlation matrix on the training numerics provides a compact readout of redundancy and economic co-movement. Key entries printed by the code:

- adj_pricedividend_ttmadj_price with dividend_ttm: 0.524, consistent with higher prices for firms with sustained payout capacity.
- unempten_yearunemp with ten_year: −0.660, a strong negative association that mirrors the typical growth-rate and rate-cycle dynamics.
- cpicpi_yoycpi with cpi_yoy: 0.616, because the level and its annual rate share trend.
- cpi_momten_year_momcpi_mom with ten_year_mom: 0.280, a modest positive link across short-horizon moves.
- dividenddividend_ttmdividend with dividend_ttm: 0.288, as one month's payout contributes a small part to the trailing total.

Across predictors we also see high correlations among related variants, such as CPI level with its changes and unemployment with its change. That informs feature selection: we can keep the change versions, drop the levels, or use regularization to control variance if we include both. The weak association between adj_price and macro change variables suggests that returns rather than price levels may be a better modeling target in the next stage, with dividend yield constructed from dividend_ttm and price.

Actionable takeaways for preprocessing. Handle CPI gaps with a consistent rule such as within-metro forward-fill after the first valid month or simple interpolation. Standardize all numeric columns to zero mean and unit variance, especially where levels and changes coexist. Consider clipping extreme price and

dividend-TTM outliers or modeling log-returns with yield as a separate predictor. Include metro or company effects to capture baseline differences documented by the boxplots. Avoid multicollinearity by preferring either levels or changes for CPI and unemployment, guided by the correlation heatmap.

Together, these results show that the pipeline has produced a clean monthly panel, the chronological split respects temporal order and yields clear windows, and the EDA has surfaced exactly where to impute, scale, and simplify features before modeling.

3. EDA insight gained from the result

The exploratory data analysis gave me a much deeper understanding of both the dataset itself and how it connects to the research problem we defined in Week 1. The merged monthly panel combines three different perspectives of data: firm-level variables such as adjusted price, dividend, and trailing twelve-month (TTM) dividend; metro-level variables such as CPI and unemployment; and one national variable, the 10-year Treasury yield. Having all of these aligned to the same month-end dates is very valuable, because it allows us to study the link between company performance and both local and national economic conditions at the same time.

One important insight is that the dataset is structured in a way that clearly supports our research question. In Week 1, we defined the problem as understanding the structural exposures of public real estate companies to macroeconomic and regional conditions. The EDA confirmed that the dataset has the right shape to address this. Each observation ties one company to one metro area and then links it to the relevant macroeconomic series. This makes it possible to test whether some firms are more sensitive to unemployment shocks, to inflation, or to changes in interest rates, depending on where they are located and how they operate.

The correlation analysis gave several useful signals. Stock prices show a moderate positive correlation of 0.52 with dividend TTM. This suggests that firms with stronger payout capacity are rewarded with higher valuations, which is consistent with basic finance theory. Unemployment has a strong negative correlation with the 10-year Treasury yield (−0.66). This reflects well-known macroeconomic dynamics: when the economy slows, unemployment rises and interest rates fall, while in expansions the opposite happens. CPI and its year-over-year change are highly correlated (0.62), which is not surprising since both are based on

the same underlying series. This finding, however, tells us to be careful not to include both measures without some adjustment, because they may carry redundant information. Finally, the weak contemporaneous link between stock price and CPI changes suggests that inflation does not show up immediately in firm-level prices, but instead may affect them indirectly through financing costs or with a lag. This is an important insight for model design, because it means we should consider creating lagged variables or interaction terms.

Another insight from the EDA is the balance of coverage across firms. The training set includes 538 rows, with each of the six firms (BXP, ELME, EQR, REXR, SLG, TRNO) contributing about 89–90 rows, and all covering roughly the same time span from 2014 through mid-2022. This balance is encouraging, because it reduces the risk of biased results that come from one firm dominating the sample. It also means we can include firm or metro fixed effects in later models without worrying about severe imbalance.

Finally, the descriptive statistics provided an overview of the scale of each variable. For example, the mean adjusted price in the training set is 54.9, mean dividend is 0.41, mean dividend TTM is 4.58, mean unemployment is 5.17, and mean 10-year Treasury yield is 1.98. Change variables such as CPI month-over-month and unemployment month-over-month are centered close to zero. This confirmed that differencing stabilized the change features, while level variables remain on different scales. The implication is that some form of standardization or scaling will be needed before modeling, especially if we use algorithms that are sensitive to scale.

4. Data Problems and Recommendations

The EDA tells us that the dataset is both promising and challenging. On the opportunity side, it provides enough coverage to study how real estate companies respond to economic changes over time and across regions. With multiple companies, multiple metros, and multiple macro variables, we can explore cross-sectional differences as well as temporal patterns. The long time span of macro variables also opens the door to examining multiple economic regimes, such as low-rate versus high-rate environments.

On the challenge side, the EDA revealed several limitations. Missing values are one of the main issues. CPI and its change variables have only 351 non-null rows in the training set, out of 538. That means about one third of the data is missing, and the missingness is concentrated in the early years of the series. Unemployment has 14 missing values as well. If we do not handle this carefully, the effective sample size will shrink and models may be unstable. Another limitation is the strong correlations among related

variables. For example, CPI and CPI year-over-year are closely tied, and unemployment is strongly linked to its month-over-month change. Including both levels and changes creates multicollinearity, which can inflate variances and reduce interpretability. Dividend data presents another challenge, since many months have zero dividends, which makes the distribution very skewed. Finally, while the dataset covers six firms, the cross-sectional dimension is still limited, which constrains how general our findings will be.

4.1 Data problems identified

Based on the EDA, the following specific data problems were identified:

Missing values: CPI and its derived variables are missing in about one third of the training set rows. Unemployment has 14 missing rows.

Skewed distributions: Price and dividend totals are heavily right-skewed, with many zero entries in dividends.

Multicollinearity: CPI levels and changes are highly correlated, and unemployment also overlaps strongly with its changes.

Scale differences: Firm-level, metro-level, and national variables are on very different numerical scales.

Limited cross-section: Only six firms are included, which limits generalization.

4.2 Recommendations for preprocessing

The EDA points to several preprocessing steps that can make the dataset more usable:

Handling missing values. For CPI, a reasonable approach is forward-fill or backward-fill after the first valid observation in each metro, combined with recomputing changes from the filled levels. This will preserve continuity without creating artificial jumps. It is also recommended to create a binary flag to indicate which rows were imputed, so the model can learn about uncertainty.

Standardizing variables. Because variables like stock prices, CPI, and interest rates are on very different scales, z-score standardization (subtracting the mean and dividing by the standard deviation) will put

them on a comparable footing. This helps avoid dominance by large-scale variables in models like regression or distance-based algorithms.

Transforming skewed variables. For prices and dividend totals, applying a log transform or winsorization can reduce the effect of extreme values. Alternatively, dividend yield (dividend TTM divided by price) can be used as a more interpretable and stable feature than raw dividends.

Reducing redundancy. Since CPI and its changes are highly correlated, we should select either levels or changes, not both. Similarly, for unemployment we should decide whether to use the level or the change. This will reduce multicollinearity and improve interpretability.

Using fixed effects. Because boxplots showed differences across metros, adding company or metro fixed effects will help absorb baseline differences. This allows the model to focus on within-metro or within-company changes over time.