Dataset ingestion and exploration

1. What is the target variable and why?

Our outcome is next-month total shareholder return (TSR) for each firm, measured monthly. In plain terms, TSR at month t+1 is the combined effect of the stock's price change over the next month plus any cash dividend received in that same next month, all scaled by the price at the end of month t. We record it on one line as TSR_{t+1} = (P_{t+1} − P_t + D_{t+1}) / P_t.

We choose next-month TSR because it matches our predictive question and our data cadence. All predictors we observe—city CPI, city unemployment, and the U.S. 10-year Treasury yield—are naturally monthly. Using the outcome one month ahead keeps time order clean: information known at month t is used to explain what shareholders experience in month t+1. This design reflects the professor's guidance to keep the question focused and practical, with lagged indicators and month-end alignment to avoid look-ahead bias.

TSR is also the right measure for real-estate equities, where dividends are an important part of total payoff. Price-only returns can mislead in months when a firm pays a distribution; TSR consolidates price and income in the same unit (percentage points) and therefore preserves what matters to investors. Because TSR is scale-free, it is directly comparable across firms with different price levels and payout policies, which supports our plan to rank six names each month without extra adjustments.

The one-company-per-metro mapping keeps the geographic link transparent at this stage. Each outcome is attached to one "home" metro, acknowledging that metro economies do not move in lockstep. This mapping is deliberately simple and aligns.

Measurement follows a few operational rules so the target is reproducible. Prices are aligned to the last trading day of each month; when an exact month-end print is unavailable due to a holiday or a data gap, we use the closest trading day within the month. We use adjusted prices so that splits and similar corporate actions do not create artificial jumps in the target. Special cash distributions that data vendors classify as dividends are included; non-cash events such as spin-offs are handled through the adjusted price series. Months with missing target

values are rare; at most one consecutive month may be forward-filled in a clearly contiguous period, otherwise the observation is dropped. This light-touch policy keeps the panel clean without over-engineering the dataset.

We compute arithmetic returns rather than log returns to keep interpretation straightforward ("basis points of return next month"). Arithmetic TSR aligns naturally with how investors think about portfolio tilts and with the simple accuracy measures we will report later. While log returns can be convenient for some statistical properties, they are not necessary for our purpose and would make the write-up less accessible.

The target also respects seasonal and institutional realities of this sector. Real-estate companies and REITs often pay on regular cycles (for example, quarterly). A monthly TSR will therefore show stretches with D_{t+1}=0 and months with a positive distribution; this is a feature, not a bug, because it matches what shareholders actually receive. Any months affected by unusual corporate events will be footnoted so that readers can distinguish genuine market moves from mechanical accounting effects.

Several alternatives were considered and set aside for the initial pass. Price-only returns understate the experience of income-paying equities. A binary up/down label would simplify modeling but would discard magnitude information that we need for ranking six names. Abnormal returns (TSR minus a benchmark) introduce an additional modeling layer and benchmark sensitivity before we have established basic signal value. Multi-month TSR reduces the number of usable observations and blurs the timing between local conditions and realized outcomes, which runs against our goal of a crisp, one-step-ahead forecast.

Finally, the target is well-connected to the feature groups we will analyze, without turning this section into a methods manual. Local CPI and local unemployment summarize local demand; the 10-year yield summarizes financing conditions; dividend yield and lagged TSR summarize company trends; and month effects capture seasonal patterns commonly seen in real estate and capital markets. By defining TSR carefully and placing it one month ahead, we create a disciplined bridge between these observable facts at time t and what shareholders experience at time t+1.

In short, next-month TSR is a focused, comparable, and decision-relevant outcome. It reflects the shareholder's actual payoff, aligns with our monthly indicators, respects time ordering, and supports transparent evaluation. Most importantly, it keeps the geography-first design clear so we can test the core idea: metro-level economic conditions and the rate backdrop can help explain what happens to shareholder returns in the month that follows.

2. What are the predictors and why?

Because our target variable is TSR at time t+1, all predictors must be observable at time t and possess clear economic or financial logic. This ensures strict temporal consistency and avoids forward-looking bias. Our team identified the following predictors: CPI, dividends, 10-year Treasury yield, lagged TSR, unemployment rate, and region. These six predictors fall into three broad categories: CPI, Treasury yields, and unemployment rate represent macroeconomic indicators; dividends correspond to corporate metrics; and region reflects geographical characteristics. Furthermore, since the primary operating regions of the companies we selected are largely concentrated around their headquarters or span broader, unclassifiable areas, we believe headquarters location constitutes a more significant regional factor. Therefore, we will use the company headquarters location to define regions.

I. Macroeconomic Indicators

CPI reflects the overall price level of local goods and services, serving as a key inflation gauge. In the real estate market, rising CPI typically yields two outcomes. First, it may drive up rental rates, thereby increasing rental income for real estate companies. Second, it also elevates operational costs—such as property management, labor, or insurance—potentially compressing profit margins. If CPI rises steadily while interest rates remain relatively stable, real estate companies may benefit from higher rents, boosting Total Shareholder Return (TSR). However, when CPI increases alongside rapidly rising interest rates, financing costs escalate and valuations compress, negatively impacting TSR. Therefore, our analysis considers not only the absolute level of CPI but also its year-over-year and month-over-month changes as predictive factors to capture directional signals of inflation trends.

The 10-year Treasury yield serves as a widely used risk-free rate benchmark in capital markets, directly influencing real estate valuations and capitalization rates. Rising long-term rates prompt investors to discount future cash flows more rigorously, pushing up capitalization rates and depressing property valuations, thereby pressuring TSR. Conversely, falling rates benefit both valuations and share prices, potentially driving TSR upward. Moreover, rapid yield fluctuations themselves trigger market capital reallocation and influence investor sentiment. Therefore, we incorporate both the interest rate level and its monthly change into predictive factors. Notably, interest rates and CPI often exhibit a strong correlation, potentially introducing multicollinearity issues. We will control for this through regularization or differential variables in subsequent modeling.

The unemployment rate measures regional economic vitality. For real estate companies, higher unemployment typically implies reduced household purchasing power, lower occupancy rates, and insufficient corporate leasing demand. This negatively impacts cash flow and TSR. Conversely, declining unemployment boosts leasing demand, helping maintain stable cash income and valuation levels. Our team employs the unemployment rate level to reflect overall regional economic health, while its month-over-month change captures short-term shocks. Given differences in economic structures across cities, identical unemployment rate shifts may impact real estate companies differently in various locations. Therefore, the interaction between the unemployment rate and city dummy variables will be part of our further examination in subsequent research.

II. Firm-Level Metrics

Dividend yield directly influences shareholder returns for real estate companies and constitutes a significant component of returns. We define it as the trailing twelve-month (TTM) cumulative dividends divided by the current month's stock price, rather than the actual dividend paid the following month. This approach not only enhances data forward-lookingness but also prevents “leaking” future information into predictive models. The level of dividend yield represents both the direct cash distribution and reflects the market's pricing of corporate risk. A high dividend yield may indicate stable cash returns or suggest market reservations about the company's prospects. In either case, dividend yield is an indispensable factor in predicting TSR.

In data processing, we employ the TTM method to smooth dividend fluctuations and flag exceptional dividend events to prevent model noise.

Lagged TSR embodies momentum. Financial markets exhibit pervasive momentum and reversal effects, where past performance may persist or correct under specific conditions. Using TSR at time t as a predictor captures market inertia and investor sentiment. For instance, when a company's stock price recently demonstrates strong performance, investors may anticipate this trend continuing, thereby driving TSR upward in the subsequent period. Conversely, if TSR reaches an unusually high peak, it may be corrected by the market in subsequent periods. The introduction of lagged TSR not only enhances the accuracy of our short-term forecasts but also provides the model with a proxy variable for gauging market sentiment. In our analysis, we will focus on examining the differential performance of the momentum effect under different macroeconomic conditions.

III. Regional Characteristics

Region will be treated as a dummy variable to control for unobservable effects stemming from a company's headquarters city. We selected six companies corresponding to the following cities: BXP—Boston, ELME—Washington, EQR—Chicago, REXR—Los Angeles, SLG—New York, TRNO—Miami. These dummy variables capture city-level differences that are difficult to quantify, including legal and tax policies, economic structure, and capital market environments. For instance, companies headquartered in New York may enjoy higher valuations in capital markets due to liquidity advantages, while those in Washington may exhibit distinct risk profiles owing to higher concentrations of government tenants. Introducing dummy variables effectively controls for these structural differences, thereby avoiding bias from omitted variables. Furthermore, dummy variables can interact with macroeconomic indicators (such as CPI or unemployment rates) to reveal how identical conditions produce differing outcomes across cities. During modeling, given limited sample sizes, dummy variables may introduce additional degrees of freedom, necessitating robustness through regularization methods or panel fixed-effects models.

IV. Summary

The predictive factor system adopted in this project includes: CPI, 10-year government bond yield, unemployment rate, dividend yield, lagged TSR, and the city of the corporate headquarters. We believe this combination comprehensively covers macroeconomic environments, company-level characteristics, and regional variations, fully explaining the primary drivers influencing TSR. The research objective is to forecast returns over the next 3 to 6 months.

3) Exploration of the dataset: variables, types, and basic stats

The analysis relies on a compact monthly panel assembled from equity pricing/dividends, metro-level CPI and unemployment, and the U.S. 10-year Treasury series. After cleaning and joins, the final table contains 769 rows and 13 columns. Each row is a company–month observation on a month-end time grid from December 2014 through August 2025. The primary key is (company, date) and there are no duplicated keys after aggregation. Companies are mapped one-to-one to metros solely to align local macro series; this mapping is fixed across the sample and used only for data joining. All timestamps are normalized to the last trading day of each month to maintain a consistent index across sources that report on different calendars.

Column typing is standardized for reliability and ease of transformation. The dataset has one datetime field (date), two categorical identifiers (company, metro), and ten numeric columns stored as float. Numeric variables include adjusted price at month-end, the within-month cash dividend, and a trailing-twelve-month dividend sum that is precomputed for later use. Macro inputs comprise the CPI index level for the assigned metro with its month-over-month and year-over-year rates, the metro unemployment rate with its month-over-month change (expressed as a percentage-point delta), and the 10-year Treasury yield with its month-over-month change. All percentage and rate fields are retained as numeric floats rather than strings so rolling windows and arithmetic are well-defined.

Basic quality checks precede the merge. Raw files are coerced to the expected types, date parsing is explicit, duplicate rows are dropped, and out-of-schema columns are excluded. After merging, we verify strictly increasing monthly sequences within each company, confirm the

uniqueness of the (company, date) key, and run a quick completeness pass. Because macro histories extend further back than some firm histories, a small number of CPI-derived fields at the earliest firm months remain missing by design; these NA values are left untouched to avoid implicit imputation. Unemployment and Treasury series align cleanly on the monthly grid after resampling to month-end. Spot checks of units ensure that price and dividends are in USD, CPI-rate fields are decimals (e.g., 0.02 for 2%), unemployment is numeric, and Treasury yields and their changes are in percentage points.

In sum, the dataset is deliberately lean: one canonical time index, two identifiers, and a small set of well-typed numeric fields. The shape (769 × 13), fixed company–metro mapping, consistent month-end timing, and explicit handling of missing edges make the table straightforward to describe, reproduce, and extend without introducing additional complexity.