***Methods:***

Building on last week’s EDA, we refined our preprocessing by handling missing data with a meaningful strategy: dropping features with high missingness > 30%, imputing moderate missingness with Iterative Imputer (MICE), and imputing low missingness with mean or mode . A key change involved addressing severe class imbalance using techniques like SMOTE and scale\_pos\_weight for XGBoost. For feature engineering, we introduced lagged variables (e.g., COVID\_Period\_Lag1) and Buffett-relevant ratios such as Profit\_Per\_Leverage. These updates aim to improve both predictive performance and model interpretability in identifying Buffett-style stock picks.

**What methods did you use to perform preprocessing? Why did you make the choices that you did?**

The preprocessing involved handling missing values, outliers, and categorical variables to prepare the dataset for modeling. Columns with more than 30% missing values, such as Inv Gr and Div Yld, were dropped to avoid introducing bias. Moderate missingness in important features like P/FCF and ROE was imputed using Iterative Imputer (MICE) to preserve valuable information, while low missingness was handled with mean imputation for numeric columns and mode imputation for categorical ones. Outliers in highly skewed features, such as P/E and Debt/Equity, were addressed using Winsorization and log transformations to reduce their impact on the model. Rare categories in the Sector column were grouped into "Other" to simplify the dataset and prevent overfitting.

These choices were made to ensure data completeness, reduce noise, and improve model interpretability. Dropping highly correlated features like PPI YOY Index and Debt/EBITDA minimized redundancy, while feature engineering added meaningful variables like COVID\_Period and Profit\_Per\_Leverage to capture temporal and financial insights. Encoding techniques, such as one-hot encoding for Sector and label encoding for binary variables, ensured compatibility with machine learning algorithms. These steps aimed to create a clean, balanced, and informative dataset for predictive modeling.

**What methods did you use to perform feature engineering? Why did you make the choices that you did?**

The feature engineering process focused on creating new variables and transforming existing ones to enhance the dataset's predictive power and interpretability. Key features added include COVID\_Period, a binary flag capturing the pandemic's impact on stock performance, and Profit\_Per\_Leverage, which measures return on equity relative to debt-to-equity, reflecting Buffett's preference for financially efficient companies. Additionally, the PE\_Category feature was created by binning the P/E ratio into interpretable categories (e.g., "Low," "Good," "High") to simplify a noisy continuous variable and align with Buffett's value-investing principles. Rare categories in the Sector column were grouped into "Other" to reduce noise and prevent overfitting.

These choices were made to capture meaningful patterns in the data while improving model interpretability. For example, COVID\_Period accounts for a significant temporal shock, while Profit\_Per\_Leverage highlights a key financial efficiency metric. Binning the P/E ratio into categories reduces the impact of extreme outliers and makes the feature more interpretable for machine learning models. Grouping rare sectors ensures that the model focuses on dominant trends rather than overfitting to sparse categories. These engineered features aim to align the dataset with Buffett's investment philosophy and improve the model's ability to predict his stock selections.

combined. Stock data and purchase activity merged on Ticker Quarter and Year. Stock data and economic data merged on Quarter and Year.

**Cleaning and Handling Missing Values**

1. **High Missingness (>30%)**: Columns like Inv Gr and Div Yld were dropped due to their high percentage of missing values (40.12% and 33.07%, respectively). Retaining these columns would introduce bias or require excessive imputation, which could distort the dataset.
   * **Justification**: Dropping these columns ensures data integrity and avoids introducing noise from unreliable imputations.
2. **Moderate Missingness (10%-30%)**: Important features such as P/FCF, ROE, and Debt/Equity were imputed using Iterative Imputer (MICE). This method leverages relationships between variables to provide more accurate imputations.
   * **Justification**: These features are critical for modeling Buffett's investment decisions, and MICE preserves their predictive value.
3. **Low Missingness (<10%)**: Numeric columns with low missingness were imputed using the mean, while categorical columns like Sector were imputed with the mode or filled with "Unknown". Additionally, Market Value (%) was imputed using the median within each sector.
   * **Justification**: Mean and mode imputations are simple and effective for low-risk missingness, while sector-wise median imputation ensures consistency within groups.
4. **Duplicates**: Duplicate rows were identified and removed to maintain data integrity.
   * **Justification**: Removing duplicates prevents overrepresentation of certain data points, ensuring unbiased modeling.

**Collapsing Variables**

* Rare categories in the Sector column (those with <3% frequency) were grouped into an "Other" category.
  + **Justification**: This reduces noise and prevents overfitting by simplifying the categorical variable while retaining meaningful distinctions.

**Mathematical Transformations**

1. **Winsorization**: Features with extreme outliers (e.g., P/E, Debt/Equity, Profit Margin) were capped at the 1st and 99th percentiles.
   * **Justification**: Winsorization reduces the influence of extreme values without discarding data, improving model stability.
2. **Log Transformation**: Skewed features were log-transformed using log1p to normalize their distributions.
   * **Justification**: Log transformation reduces skewness, making the data more suitable for machine learning algorithms that assume normality.

**Encoding**

1. **Label Encoding**: Binary variables like COVID\_Period were label-encoded.
   * **Justification**: Label encoding is efficient for binary variables and ensures compatibility with machine learning models.
2. **One-Hot Encoding**: Categorical variables like Sector were one-hot encoded.
   * **Justification**: One-hot encoding allows the model to interpret categorical variables numerically without introducing ordinal relationships.

**Feature Engineering**

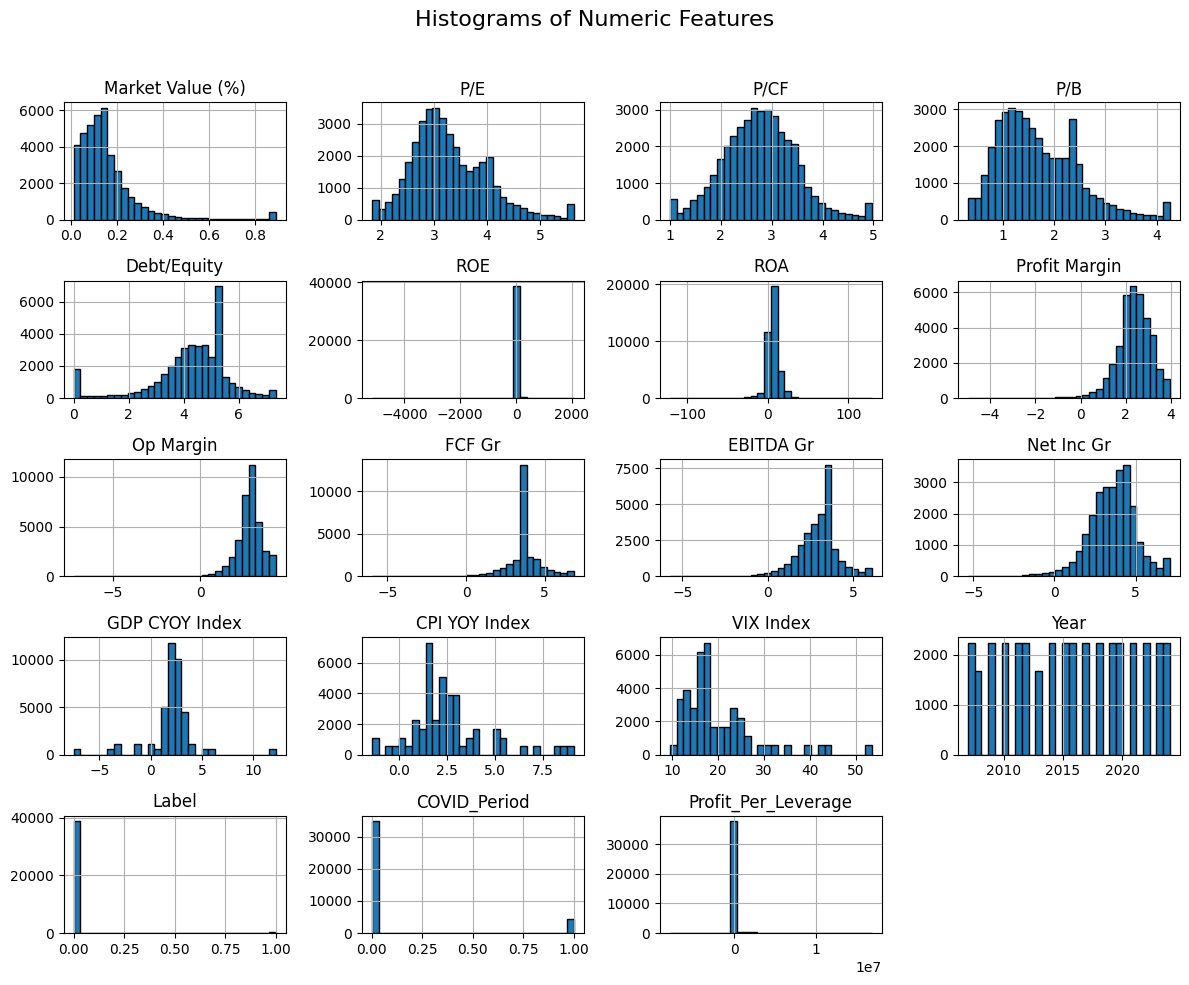
1. **COVID\_Period**: A binary flag was created to capture the impact of the COVID-19 pandemic on stock performance.
   * **Justification**: This feature accounts for a significant temporal shock that affected different sectors differently.
2. **Profit\_Per\_Leverage**: Calculated as ROE / Debt-to-Equity to measure financial efficiency.
   * **Justification**: This metric aligns with Buffett's preference for companies with strong returns and manageable debt levels.
3. **PE\_Category**: The P/E ratio was binned into categories (Low, Good, High, etc.) based on Buffett-style logic.
   * **Justification**: Binning simplifies a noisy continuous variable into interpretable categories, improving model interpretability.

**Visualizations**

1. **Histograms**: Visualized the distribution of numeric features before and after transformations to confirm reduced skewness.

A screenshot of a graph

AI-generated content may be incorrect.



1. **Boxplots**: Used to identify and confirm the reduction of outliers after Winsorization and log transformations.

A graph with different colored boxes

AI-generated content may be incorrect.

A graph with different colored boxes

AI-generated content may be incorrect.