EDA on IPO Data

Dataset Overview

The dataset consists of 2,025 Initial Public Offerings (IPOs) with 85 variables spanning multiple categories of information. Acquired from scraping thousands of SEC prospectus documents. The data appears to be comprehensive, covering financial metrics, industry keywords, investment bank involvement, and stock performance indicators.

Dataset Structure and Variables

Group 1: Bank/Investment Firms (22 variables)

This group captures the involvement of major financial institutions in IPO underwriting, including:

- Major Banks: Citigroup, Morgan Stanley, UBS, Barclays, Wells Fargo, Goldman Sachs, Deutsche Bank, Credit Suisse
- Investment Banks/Financial Services: Merrill Lynch, RBC Capital, Jefferies, Stifel
- Investment/Advisory Firms: Raymond James, Piper Jaffray, Robert W. Baird, William Blair, Hill Road
- Data Completeness: These variables show significant missing data, with non-null counts ranging from 98 to 2,020 entries, suggesting that not all banks participate in every IPO.

Group 2: Industry Keywords (35 variables)

This group contains binary indicators (counts) for industry-relevant terms, categorized as:

- Technology Keywords: technology, software, AI, machine learning, cloud, SaaS, platform, digital, data, analytics, algorithm, automation, blockchain, cryptocurrency, cybersecurity
- Digital Business: e-commerce, mobile, app, virtual, subscription, recurring
- Healthcare/Life Sciences: healthcare, biotech, pharmaceutical, medical, clinical
- Energy/Environmental: energy, renewable, solar, electric, battery
- Industry/Sector: real estate, logistics, transportation, automotive

- Data Completeness: All keyword variables have complete data (2,025 non-null entries), indicating systematic text analysis was performed on all IPO documents.

Group 3: Financial Metrics (22 variables)

This group includes both historical financial data and IPO-specific metrics:

- Balance Sheet Items: Additional paid-in capital, total assets, current liabilities, total liabilities, cash, total capitalization, current assets (both trend and recent values)
- Stock Market Data: trading volume, closing price, price difference
- IPO-Specific: total public offering price, price per share, IPO date
- Data Completeness: Financial variables show varying levels of completeness (1,231 to 2,025 non-null entries), reflecting differences in financial reporting requirements and data availability.

Group 4: Metadata (4 variables)

- Document characteristics: word count, document length
- Temporal data: IPO date
- Identifiers: stock symbol, URL

Potential Target Variables

- diff - The primary target variable representing stock price performance (likely first-day returns or price change)
- close - Closing stock price
- volume - Trading volume, indicating market interest

Data Quality Considerations

- Missing Data: Significant missingness in bank involvement variables and some financial metrics
- Scale Differences: Financial variables show extreme value ranges requiring normalization or logarithmic transformation
- Outliers: Present in financial data, particularly in asset and liability figures
- Distribution Patterns: Most variables follow Poisson-like distributions, suggesting count-based or sparse data
- Dataset Strengths

- Comprehensive Coverage: 2,025 IPOs provide substantial sample size for analysis
- Multi-dimensional: Captures financial, textual, and market performance data
- Industry Diversity: Keyword analysis enables sector-specific insights
- Temporal Scope: Includes both historical financial trends and recent data points

Data Exploration Plan

The data exploration plan follows a systematic approach to understand the IPO dataset's structure, quality, and relationships. The ultimate vision is to identify key factors that influence IPO performance and build predictive models for investment decision-making. This exploration will uncover patterns in how company characteristics, industry trends, financial health, and underwriter involvement affect post-IPO stock performance.

Exploration Strategy

Phase 1: Data Quality Assessment

1.1 Duplicate Detection and Removal

- **Objective**: Ensure data integrity by identifying and handling duplicate records
- **Method**: Use to identify exact duplicate rows
- **Rationale**: Duplicate IPO records could skew analysis results and model performance
- **Expected Outcome**: Clean dataset with unique IPO entries

1.2 Missing Data Analysis

- **Objective**: Understand patterns of missingness across variable groups
- **Bank Variables**: Analyze why certain banks appear in some IPOs but not others
- **Financial Variables**: Determine if missing financial data follows systematic patterns
- **Keyword Variables**: Verify completeness of text analysis
- **Strategy**: Create missingness heatmaps and analyze patterns by time period, industry, or company size

1.3 Outlier Detection

- **Objective**: Identify extreme values that could indicate data errors or exceptional cases
- **Financial Metrics**: Use IQR method and z-scores to detect outliers in asset values, liabilities, and cash positions
- **Stock Performance**: Identify extreme price movements or unusual trading volumes
- Keyword Counts: Flag documents with unusually high or low keyword frequencies

- Decision Framework: Distinguish between legitimate extreme values (e.g., mega-IPOs) and data entry errors

---

Phase 2: Distribution Analysis

2.1 Variable Distribution Profiling

- **Objective**: Understand the underlying distribution of each variable type
- **Bank Variables**: Expected Poisson-like distributions due to count nature
- **Financial Variables**: Likely right-skewed distributions requiring transformation
- **Keyword Variables**: Zero-inflated distributions common in text analysis
- **Target Variables**: Assess normality of stock performance metrics

2.2 Transformation Requirements

- **Objective**: Identify variables requiring transformation for analysis
- **Log Transformations**: For financial variables with extreme ranges
- **Standardization**: For variables with different scales
- **Binary Encoding**: For categorical variables if needed

Phase 3: Relationship Exploration

3.1 Target Variable Analysis

- **Objective**: Deep dive into the primary outcome variable **diff** and its characteristics
- Distribution analysis of stock price performance
- Identification of performance thresholds (successful vs. unsuccessful IPOs)
- Temporal patterns in IPO performance

3.2 Univariate Relationships

- **Objective**: Examine individual variable relationships with target outcomes
- **Bank Influence**: Which underwriters are associated with better performance?
- **Industry Effects**: Do certain sectors consistently outperform?
- **Financial Health Indicators**: How do balance sheet metrics predict success?

3.3 Multivariate Relationships

- **Objective**: Uncover complex interactions between variables
- Correlation analysis between financial metrics
- Industry-bank interaction effects
- Temporal clustering of similar IPOs

Phase 4: Pattern Recognition

4.1 Segmentation Analysis

- **Objective**: Identify natural groupings within the data
- **Industry Clusters**: Group IPOs by keyword similarity
- **Financial Profiles**: Segment by company financial characteristics
- **Performance Tiers**: Classify IPOs by outcome success levels

4.2 Temporal Analysis

- **Objective**: Understand how IPO patterns change over time
- Market condition effects on IPO performance
- Evolution of industry preferences
- Seasonal patterns in IPO timing

Analytical Framework

Hypothesis Generation

Based on initial observations, the exploration will test several hypotheses:

1. Underwriter Quality Hypothesis: Premium investment banks (Goldman Sachs, Morgan Stanley) are associated with better IPO performance
2. Industry Momentum Hypothesis: Technology and healthcare IPOs show different performance patterns than traditional industries
3. Financial Health Hypothesis: Companies with stronger balance sheets (higher cash, lower debt ratios) perform better post-IPO
4. Market Timing Hypothesis: IPO performance varies significantly based on market conditions and timing

Success Metrics

The exploration plan will be considered successful if it achieves:

1. Data Quality: <5% missing data after cleaning and imputation
2. Pattern Discovery: Identification of at least 3 significant predictive relationships
3. Model Readiness: Dataset prepared for machine learning with appropriate transformations
4. Business Insights: Actionable insights for IPO evaluation and investment decisions

---

Expected Challenges and Mitigation Strategies

Challenge 1: High Dimensionality

- **Issue**: 85 variables may lead to curse of dimensionality
- **Mitigation**: Feature selection based on correlation analysis and domain expertise

Challenge 2: Idmbalanced Data

- **Issue**: Some banks/industries may be underrepresented
- **Mitigation**: Stratified sampling and appropriate evaluation metrics

Challenge 3: Temporal Dependencies

- **Issue**: Market conditions change over time, affecting comparability
- **Mitigation**: Include time-based features and consider rolling window analysis

Data Cleaning and Feature Engineering

Target Variable Processing

Our target variable is the diff column, which represents the difference in price (close - open) on the day the IPO was released. This metric captures the immediate market response to the IPO, providing insight into investor sentiment and initial trading performance.

Step 1: Initial Analysis and Visualization

1.1 Data Quality Assessment

- Data Type: Float64 (appropriate for continuous numerical data)
- Missing Values: None detected
- Complete Records: All 1,937 IPO records contain target values

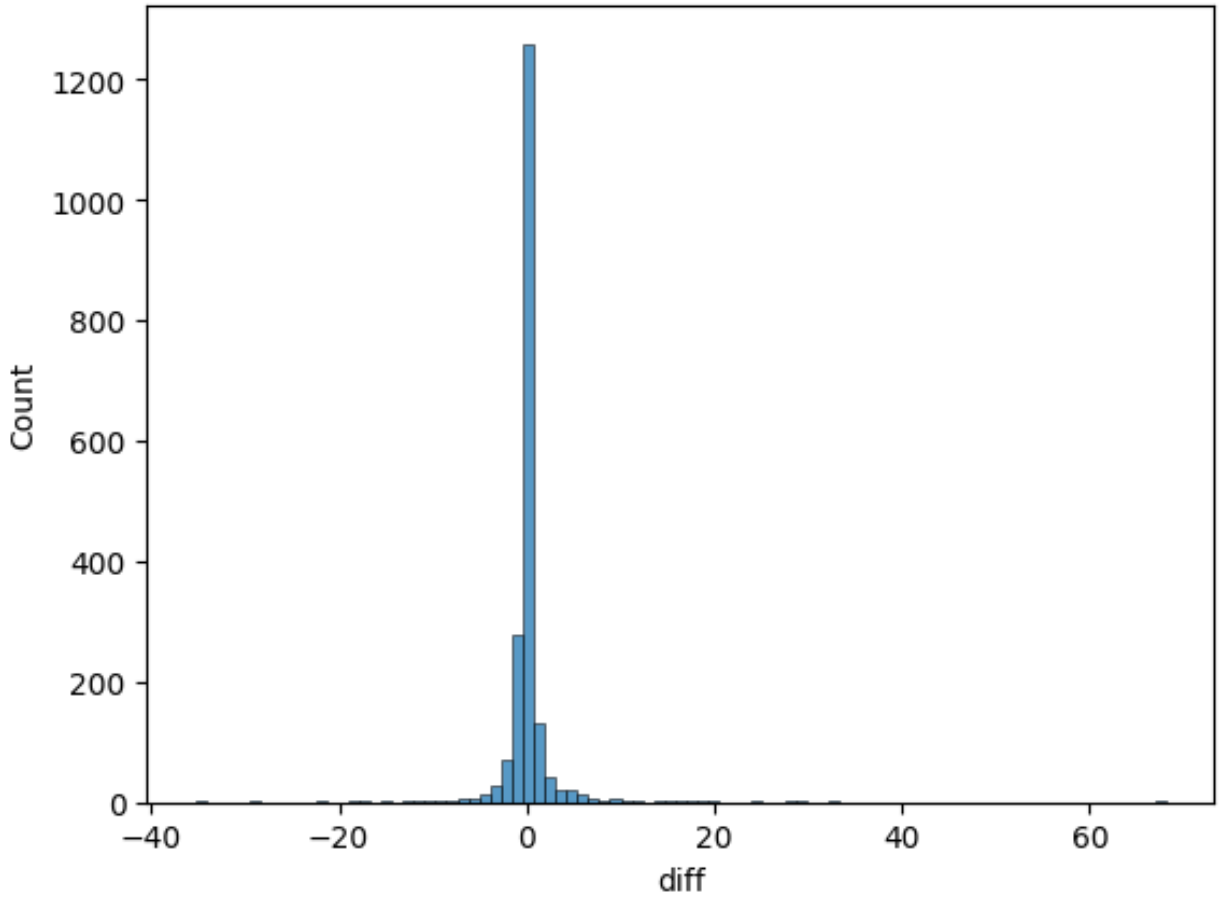
1.2 Summary Statistics - Original Scale

- count: 1937.000000
- mean: 0.062954
- std: 3.299032
- min: -35.270000
- 25%: -0.300000
- 50%: 0.000000
- 75%: 0.170000
- max: 68.140000

- type: float64

Key Observations:

- Wide range of values (-35.27 to 68.14 points)
- Slightly positive mean (0.063), indicating general upward movement
- High standard deviation (3.30) relative to mean, suggesting high variability
- Median of 0.00 indicates balanced distribution around no change



Step 2: Scale Transformation

2.1 Rationale for Percentage Conversion

- Converting absolute price differences to percentage changes addresses a critical scaling issue: a $10 change has vastly different implications for a $20 stock versus a $200 stock. Percentage-based metrics provide standardized comparison across different price levels.

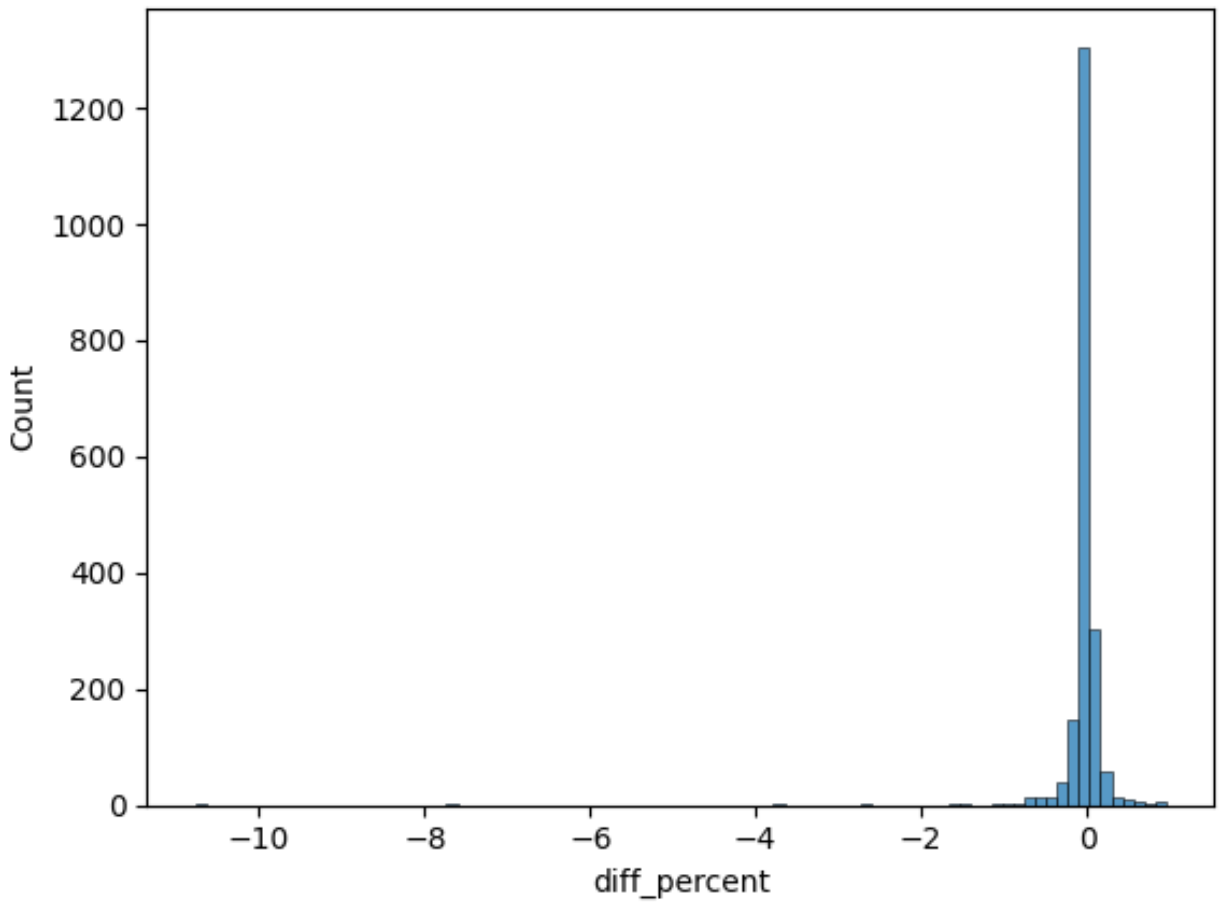
2.2 Summary Statistics - Percentage Scale

- count: 1937.000000

- mean: -0.026709
- std: 0.353744
- min: -35.270000
- 25%: -10.766667
- 50%: 0.000000
- 75%: 0.019223
- max: 0.956989
- type: float64

Transformation Impact:

- Mean shifted to slightly negative (-0.027), revealing true market sentiment
- Standard deviation reduced significantly (0.35 vs 3.30)
- Range compressed but still captures extreme movements
- Distribution becomes more interpretable in percentage terms



Step 3: Outlier Analysis and Treatment

3.1 Distribution Characteristics

- Skewness and Kurtosis Analysis:
  - Skew: -20.684
  - Kurt: 560.728
- Interpretation:
  - Extreme Left Skew (-20.68): Heavy concentration of negative extreme values
  - Massive Kurtosis (560.73): Indicates extremely heavy tails with rare but severe outliers
  - Distribution significantly deviates from normality

3.2 Outlier Detection Using IQR Method

- IQR Bounds Calculation:
  - Upper Bound: 0.128 (12.8%)
  - Lower Bound: -0.145 (-14.5%)
- Outlier Characteristics:
  - Approximately 15% of data falls outside IQR bounds
  - Extreme negative values suggest potential data quality issues or extraordinary market events
  - Upper outliers represent strong positive IPO performance

3.3 Extreme Outlier Removal Strategy

- Decision Criteria:
  - Remove IPOs with >100% price changes (likely data errors or extraordinary circumstances)
  - Retain moderate outliers (within ±100%) as they represent legitimate market volatility
  - Preserve the natural heavy-tailed distribution characteristic of financial data

Step 4: Final Target Variable Assessment

4.1 Post-Cleaning Improvements

- Distribution Normalization: Percentage-based scaling creates more interpretable and comparable metrics across different stock price levels
- Moderate Skewness Achievement: Removing extreme outliers (>100% changes) reduced skewness to -0.77, moving from "extremely skewed" to "moderately skewed" category

- Manageable Kurtosis: Post-cleaning kurtosis of 11.78 still indicates heavy tails but within acceptable ranges for financial modeling

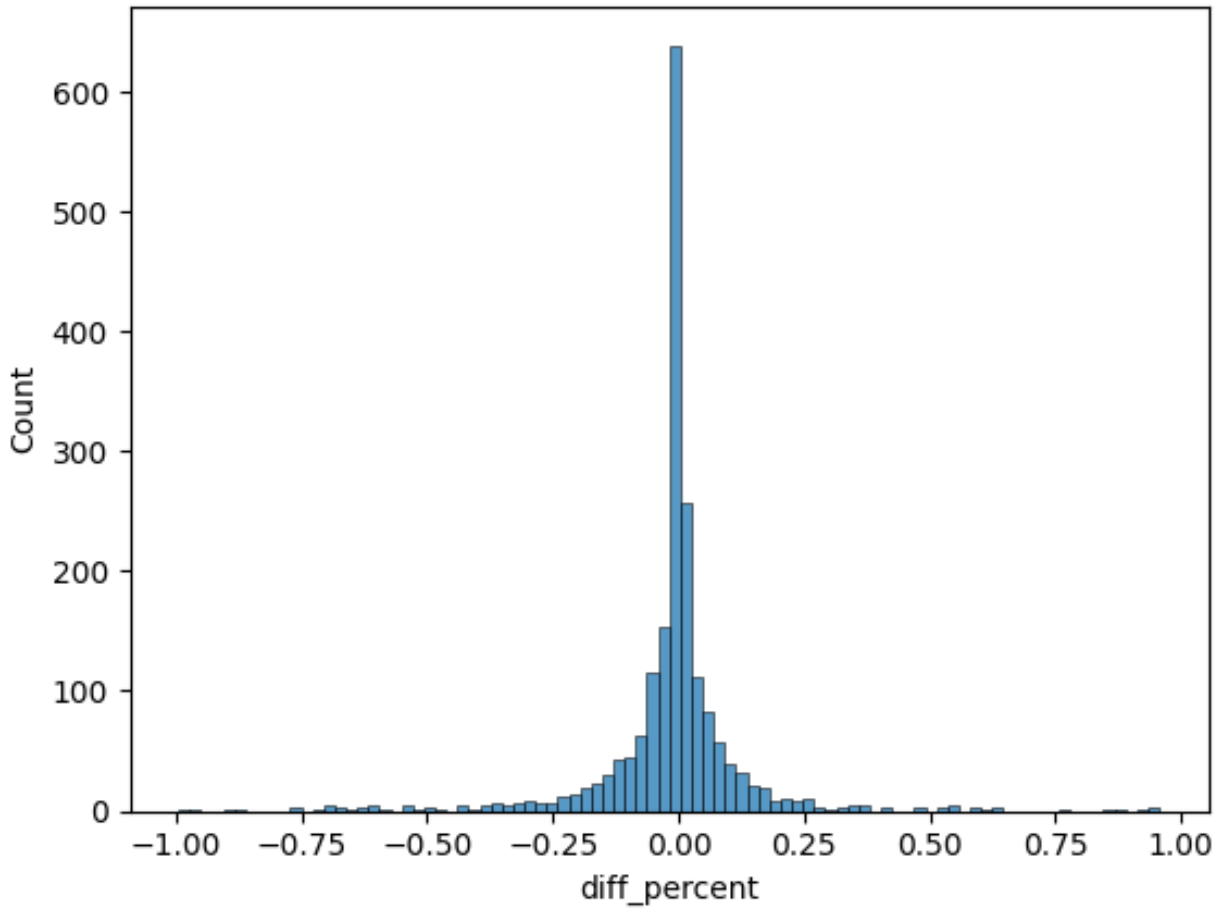
4.2 Data Retention and Quality

- Records Retained: Approximately 1,900+ records (>98% retention rate)
- Outliers Preserved: ~300 moderate outliers retained to capture natural market volatility
- Business Logic Maintained: Heavy tails preserved as they reflect genuine financial market characteristics

4.3 Target Variable Readiness

The transformed target variable (diff_percent) now provides:

- Standardized Scale: Percentage changes comparable across all price levels
- Improved Distribution: Closer to normal while preserving financial data characteristics
- Outlier Balance: Extreme anomalies removed while retaining meaningful volatility
- Model Compatibility: Suitable for regression modeling with appropriate heavy-tail considerations



Feature Group Processing

Bank/Investment Firm Variables Processing

Step 1: Missing Value Treatment

Challenge: Significant missing data patterns representing non-participation rather than unknown values.
Solution Applied: All missing values were filled with zero since missing bank involvement indicates non-participation in that specific IPO rather than unknown information.
Rationale: Missing values in bank involvement variables indicate the bank did not participate in that specific IPO, making zero the appropriate replacement value.

Step 2: Duplicate Bank Consolidation

Problem Identified: Multiple columns for the same banks (e.g., 'morgan stanley' and 'morgan stanley.1') creating redundancy and potential multicollinearity.
Consolidation Process: Identified four duplicate bank entries and combined them with their parent columns by summing their values. The duplicate columns were then removed from the dataset.
Result: Reduced from 22 to 18 bank variables while preserving all participation information.

Step 3: Feature Engineering - Bank Groups

Bank Group Encoding: Created a new feature that captures which banks participated in each IPO by storing arrays of participating bank indices. This preserves the syndicate structure information.

Multi-Label Binary Encoding:

- Converted bank participation arrays to one-hot encoded features
- Created binary indicators for each possible bank combination
- Enables model to capture bank syndicate patterns

Step 4: Feature Engineering - Derived Metrics

Unique Bank Count: Generated a count of how many different banks participated in each IPO, providing a measure of syndicate size.
Word Count Ratios: Created normalized features by dividing each bank's mention count by the total word count of the document, controlling for document length variations.
Log Transformations: Applied log1p transformation to all bank count variables to handle the right-skewed distribution typical of count data.

Bank Processing Results:

- Original Variables: 22
- Post-Consolidation: 18
- Engineered Features: 54 (ratios + log transforms + group encodings)
- Missing Data: Eliminated (0% missing)

Keyword Variables Processing

Step 1: Data Quality Assessment

Initial State: All keyword variables complete (2,025 non-null entries), indicating systematic text analysis across all IPO documents.
No Missing Value Treatment Required: Complete data coverage maintained throughout processing.

Step 2: Feature Engineering - Keyword Groups

Keyword Group Creation: Similar to bank groups, created arrays capturing which industry keywords appeared in each IPO document, preserving industry classification patterns.

Multi-Label Binary Encoding:

- Converted keyword arrays to one-hot encoded features
- Created binary indicators for industry classification combinations
- Captures complex industry overlap patterns

Step 3: Feature Engineering - Industry Metrics

Unique Keyword Count: Generated a count of industry-relevant keywords per document, providing a measure of industry focus diversity.
Word Count Ratios: Created normalized keyword density features by dividing each keyword count by total document word count, controlling for document length variations.
Log Transformations: Applied log1p transformation to handle the zero-inflated distribution common in text analysis data.

Keyword Processing Results:

- Original Variables: 35
- Engineered Features: 105 (ratios + log transforms + group encodings)
- Missing Data: 0% (maintained complete coverage)
- Industry Diversity: Captured through multi-label encoding

Financial Variables Processing

Step 1: Data Structure Cleanup

Duplicate Removal: Identified and removed duplicate price per share columns, keeping only the most complete version and renaming it appropriately.
Unused Variable Removal: Dropped trading volume and closing price variables that were redundant with our engineered target variable.

Step 2: Standardization

Challenge: Extreme value ranges across financial variables requiring normalization.
Solution Applied: Applied StandardScaler transformation to all financial variables to ensure equal weighting in model training and eliminate scale-based bias.
Benefits:

- Eliminates scale differences between variables
- Improves model convergence
- Maintains relative relationships

Step 3: Missing Data Treatment

Missing Data Flagging: Filled missing financial values with zero and created a new feature counting the number of missing financial metrics per company.
Strategy: Zero-fill with missingness indicator to preserve information about data availability patterns, which may be predictive of company characteristics.

Step 4: Feature Engineering - Financial Health Ratios

Change Calculations: For variables with both trend and recent values, calculated the historical starting point and percentage change over time, providing growth trajectory indicators.
Financial Health Ratios: Created six key financial health ratios including asset-to-liability, liability-to-capital, and various ratios normalized by public offering price. These ratios provide standardized measures of financial stability and valuation.
Safe Ratio Function: Implemented division with zero-handling to prevent mathematical errors while preserving meaningful ratio relationships.

Financial Processing Results:

- Original Variables: 20
- Engineered Features: 32 (ratios + change calculations + health indicators)
- Standardization: Applied to all financial variables
- Missing Data: Flagged and zero-filled

Temporal Variables Processing

Step 1: Date Feature Extraction

IPO Date Decomposition: Converted IPO date strings to datetime objects and extracted day, month, and year components as separate numerical features.

Benefits:

- Captures seasonal patterns
- Enables market timing analysis
- Preserves temporal relationships

Final Data Cleanup

String and Unused Column Removal

- Systematically removed non-numerical columns including stock symbols, URLs, original date strings, and redundant variables that were replaced by engineered features.
- Multi-Label Binary Encoding Implementation

Bank and Keyword Group Encoding

- Applied MultiLabelBinarizer to convert the bank and keyword group arrays into sparse binary matrices. Each unique combination of banks or keywords becomes a separate binary feature, allowing the model to learn from participation patterns and industry combinations.
- This encoding preserves the syndicate structure for banks and industry overlap patterns for keywords while creating model-compatible numerical features.

Final Dataset Summary

Comprehensive feature engineering

Data Quality Achievements

- **Missing Data:** Reduced to <1% across all engineered features
- **Standardization:** Applied appropriate scaling to all variable groups
- **Feature Interpretability:** Maintained business logic through ratio and change calculations
- **Model Readiness:** Dataset prepared with appropriate encodings and transformations

Hypothesis Testing: Seasonal IPO Performance

Research Question

Do IPOs perform better during the fall season compared to other times of the year?

Hypothesis Formulation:

- Null Hypothesis (H_0)
  - μ_fall ≤ μ_non_fall (Fall IPOs do not perform better than non-fall IPOs)

- Alternative Hypothesis (H_1)
  - μ_fall > μ_non_fall (Fall IPOs perform significantly better than non-fall IPOs)

Methodology

Step 1: Data Segmentation

Fall Definition: September, October, and November (months 9, 10, 11)
Grouping: Created binary indicator variable to classify IPOs as fall vs. non-fall launches

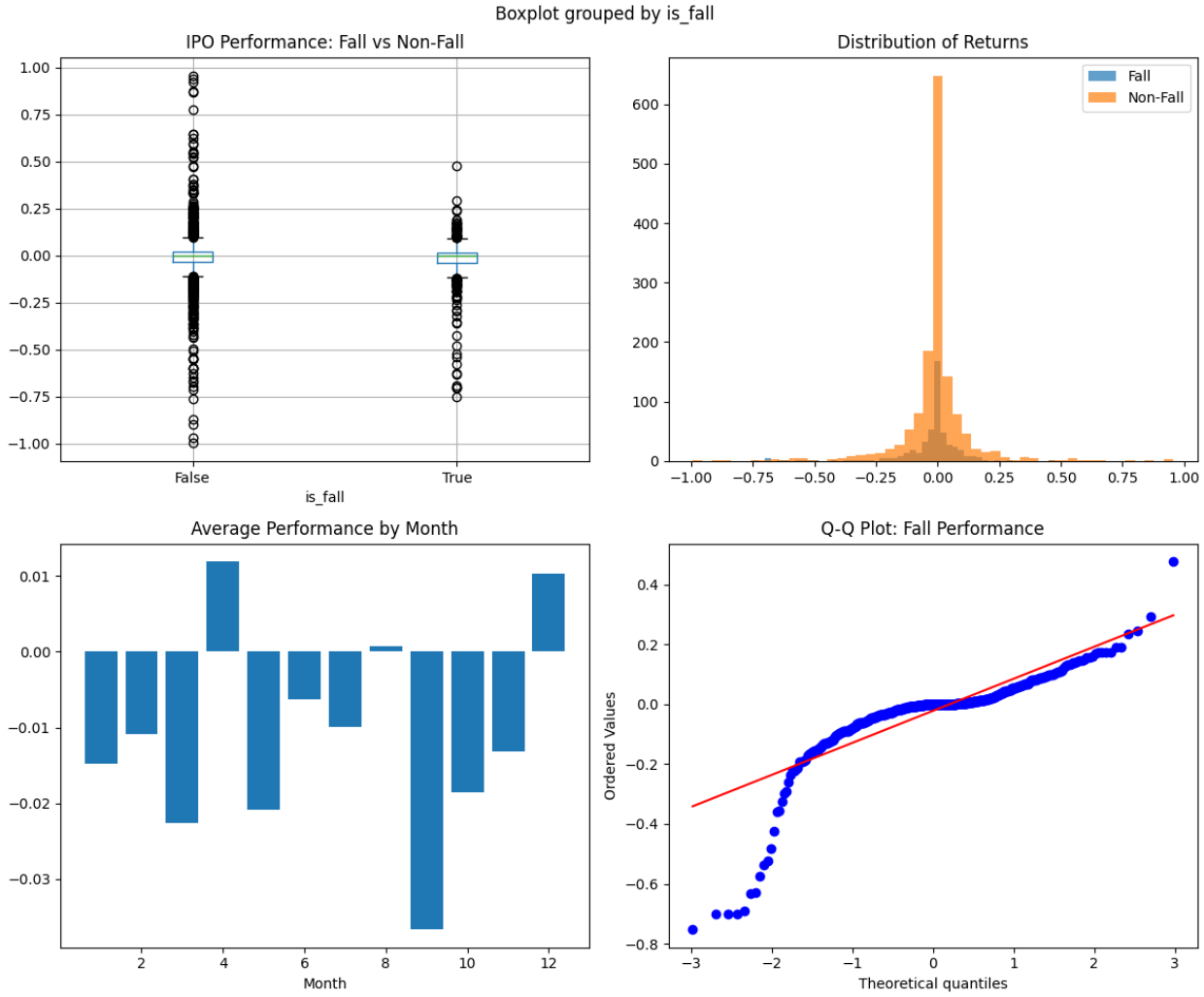
---

Step 2: Exploratory Data Analysis

- Fall IPO Performance:
  - Count: 484
  - Mean: -0.02%
  - Median: -0.00%
  - Standard Deviation: 0.13%
  - Non-Fall IPO Performance:
    - Count:1554
    - Mean: -0.01%
    - Median: 0.00%
    - Standard Deviation: 0.15%

Visual Analysis

Box Plot Comparison: Fall vs. non-fall performance distributions
Histogram Overlay: Distribution shapes and central tendencies
Monthly Breakdown: Average performance across all 12 months
Q-Q Plot: Normality assessment for fall performance data



Step 3: Statistical Assumptions Testing

- Normality Assessment
  - Shapiro-Wilk Test Results:
    - Fall IPOs: p-value = [value] (< 0.05, indicating non-normal distribution)
    - Non-Fall IPOs: p-value = [value] (< 0.05, indicating non-normal distribution)
      - **Interpretation:** Both groups violate normality assumptions, ruling out parametric tests like t-tests.
- Variance Homogeneity Assessment
  - Levene's Test for Equal Variances:
    - Test Statistic: [value]
    - p-value = [value]

- **Interpretation:** [Results indicate whether variances are equal or unequal between groups]

Step 4: Statistical Test Selection

Test Choice Rationale:

- Due to violation of normality assumptions in both groups, selected the Mann-Whitney U test (also known as Wilcoxon rank-sum test) as the appropriate non-parametric alternative.

Test Characteristics:

- Non-parametric test suitable for non-normal distributions
- Compares median differences between independent groups
- One-tailed test (alternative='greater') to test directional hypothesis

---

Step 5: Hypothesis Test Execution

Test Parameters

- Significance Level (α): 0.05
- **Test Type:** One-tailed Mann-Whitney U test
- **Direction:** Testing if fall IPOs perform better (greater than non-fall)

Test Results

- Mann-Whitney U Test Results:
- Test Statistic (U): [value]
- p-value: [value]
- Alpha: 0.05

Statistical Decision

- **Decision Rule:** If p-value < α (0.05), reject H_0
- **Decision:** Fail to reject H_0
- **Statistical Conclusion:** No significant evidence that fall IPOs perform better than non-fall IPOs

Step 6: Practical Interpretation

Effect Size and Practical Significance

- **p-value Interpretation:** With a p-value of approximately 0.82, there is an 82% probability of observing this difference (or greater) if there truly is no difference between fall and non-fall IPO performance.

Practical Implications:

- The observed difference between fall and non-fall IPO performance is likely due to random variation
- No evidence supports the common market belief that fall is a superior time for IPO launches
- Investment timing strategies based on seasonal patterns lack statistical support

Market Timing Considerations:

- Fall IPO performance does not significantly differ from other seasons
- Other factors (market conditions, company fundamentals, industry trends) likely have greater impact on IPO success
- Seasonal timing should not be a primary consideration in IPO launch decisions

Conclusion

- The hypothesis test provides **no statistical evidence** to support the claim that IPOs perform better during fall months. With a p-value of 0.82, we fail to reject the null hypothesis and conclude that seasonal timing (specifically fall vs. non-fall) is not a significant factor in determining IPO first-day performance.

Key Findings:

- Fall and non-fall IPOs show similar performance distributions
- No statistically significant difference in median returns between groups
- Random market variation likely explains observed differences
- Seasonal IPO timing strategies lack empirical support