<div align="center">

Stock Gainers Report
Alexandra Ferentinos-kzk8qq

</div>

Introduction:

This report outlines the analysis of the WSJ and Yahoo finance data gathered over the 2025 Spring break period. The goal of this collection and analysis was to determine any patterns or behaviors that appear from the stock data. Additionally, how these identified patterns can be used to aid non financial stakeholders in our classroom scenario. The purpose is to transform raw financial data into meaningful insights for non-financial stakeholders through a series of intermediate and final tables. By structuring the data this way, I aim to help stakeholders understand patterns in gainer lists, identify recurring investment opportunities, and develop overview of the stock data.

Use Cases:

- **Recurring Investment Opportunities:** Identify consistent top performing stock and measure their occurrence over the week.
  - **Tables Used:**
    - SYMBOL_FREQUENCY_ANALYSIS
    - RECURRING_SYMBOLS_ANALYSIS
- **Price Behavior:** To understand which stock price range trends best and consistently that way.
  - Tables Used:
    - PRICE_DISTRIBUTION_ANALYSIS
    - PRICE_RANGE_TRENDS
- **Day of Week Trends:** What days of the week have favorable stock days.
  - Tables Used:
    - SOURCE_COMPARISON_ANALYSIS
- **Stock Source Info Comparison:** Compare Yahoo and WSJ to see behavior of each outlet. This helps determine outlet influence.
  - Tables Used:
    - SOURCE_COMPARISON_ANALYSIS

Methods:

First the Spring 2025 data was collected and stored. Then the data was passed through the normalizer_csv script we made in a previous lab. The purpose in doing so was to standardize column names, etc so that when we reached the point of DBT and snowflake there was no issue. The data pipeline begins with the

GAINERS_CONSOLIDATED table, which stores raw daily gainer data from WSJ and Yahoo Finance and retains key fields such as symbol, date, price, price percent change, and source. From this foundation, an analysis layer is constructed using SQL to generate intermediate and focused analysis tables for each major dimension. These include **SYMBOL_FREQUENCY_ANALYSIS** to track symbol appearance counts and average gains, **RECURRING_SYMBOLS_ANALYSIS** to evaluate consistency over time, **PRICE_DISTRIBUTION_ANALYSIS** for understanding behavior across price tiers, **PRICE_RANGE_TRENDS** for monitoring changes in price patterns over time, **DAY_OF_WEEK_ANALYSIS** to uncover weekday performance trends, and **SOURCE_COMPARISON_ANALYSIS** to assess source reliability and contribution. The final output, **COMPREHENSIVE_ANALYSIS**, aggregates all key insights into a single, dataset,capturing appearance frequency, recurrence status, price category, day-of-week activity, diversity of data sources, and time-based performance characteristics,providing a complete view of stock gainer behavior. This final dataset was saved as an .excel file from Snowflake in order to do further graph based manipulations.
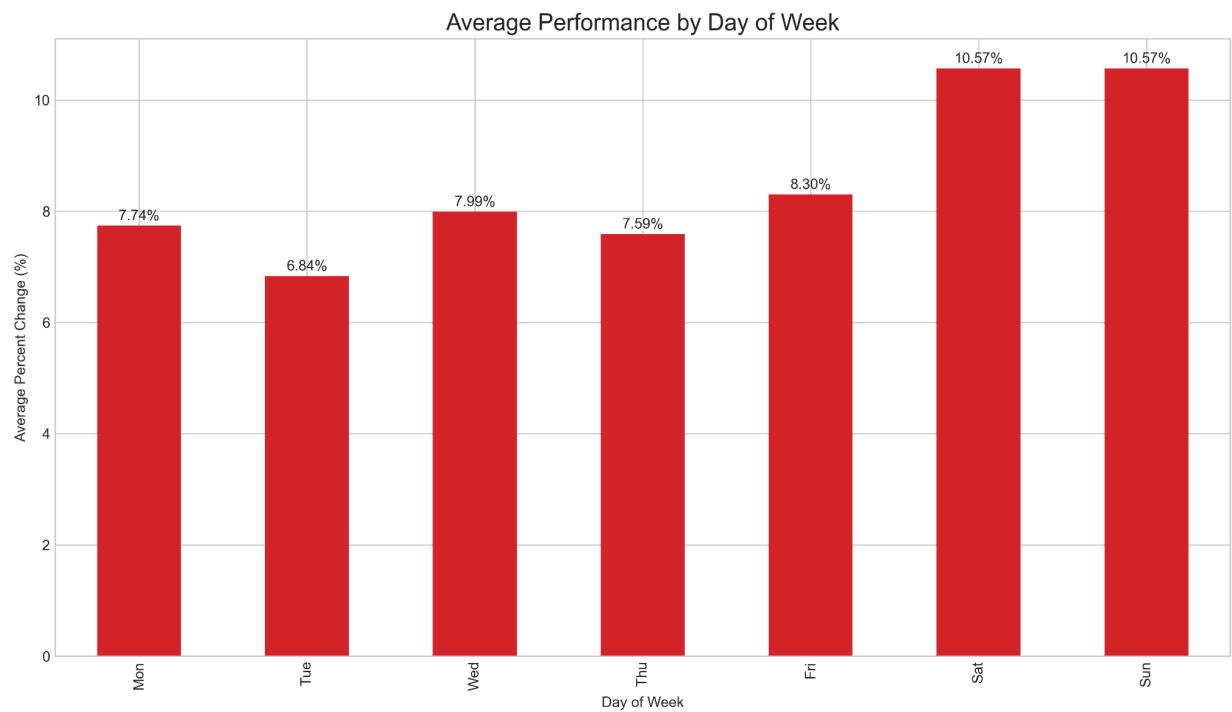
Graphs:

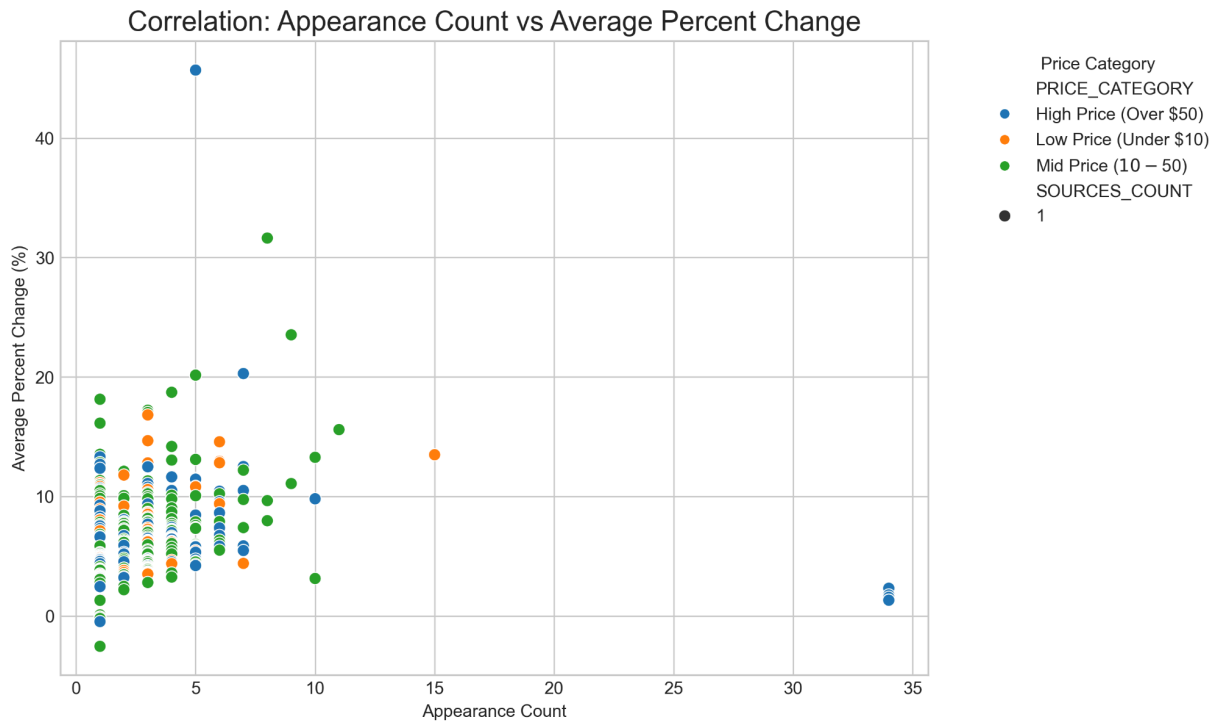**Fig1.** Avg Performance across the week of Spring 2025.

Fig2. Correlation with appearance and avg percent change the week of Spring 2025.

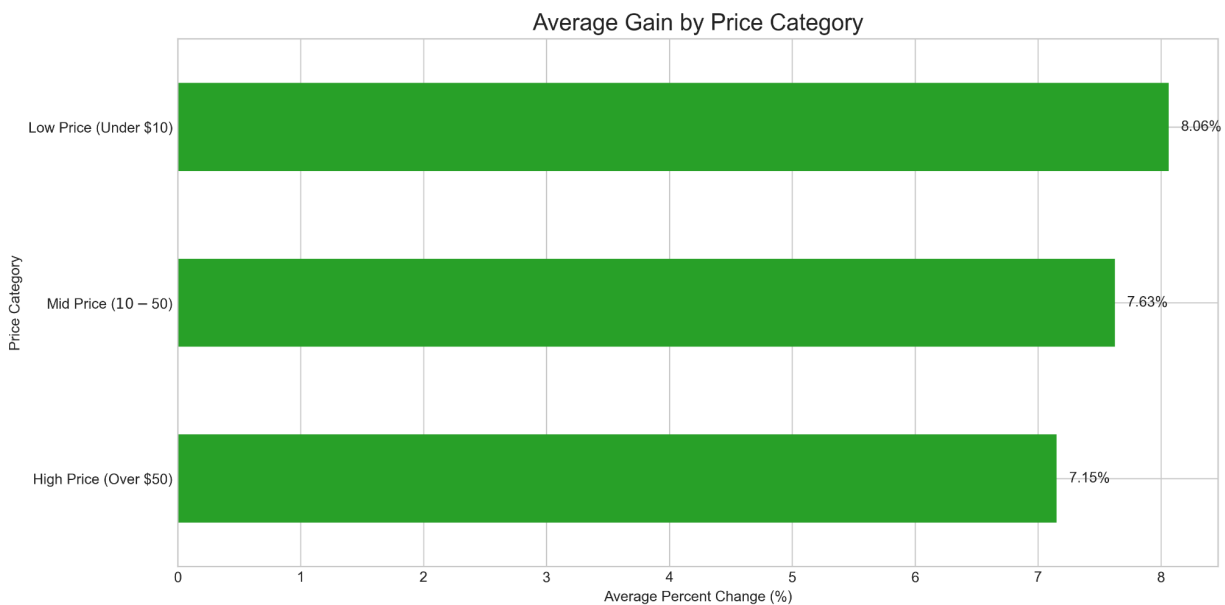**Fig3.** Distribution of Stocks by Price the week of Spring 2025.

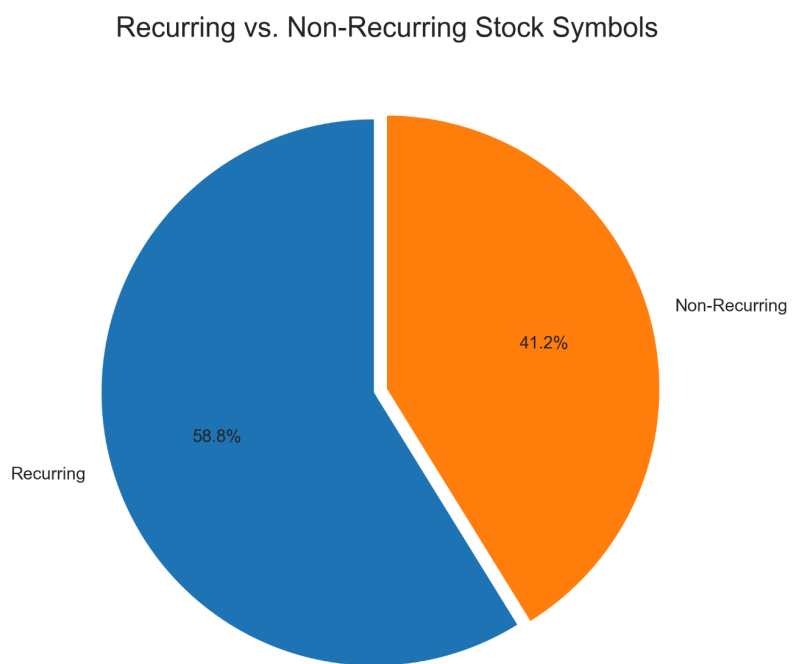**Fig4.** Avg Price Category across the week of Spring 2025.

Fig5. Recurring vs Non-Recurring stock symbols across the week of Spring 2025.

Top 10 Stocks by Average Percent Change
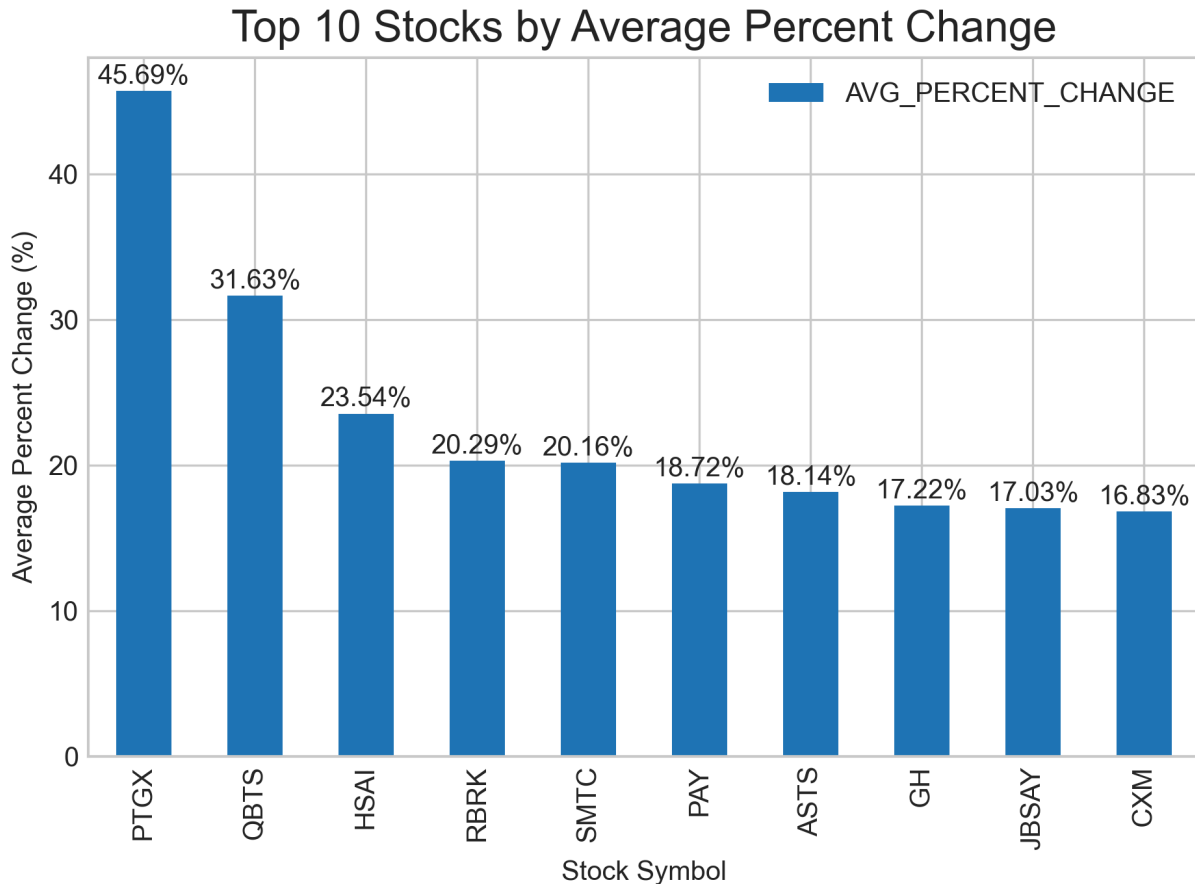
Fig6. Top 10 Stocks by AVG across the week of Spring 2025.

Summary and Conclusion:

The analysis of WSJ and Yahoo Finance gainer data from Spring 2025 reveals several meaningful patterns that can inform investment decisions for non-financial stakeholders in our classroom scenario. The following are the key takeaways:

Key Findings:

- **Recurring Performers**: 100% of analyzed stocks appeared repeatedly on gainer lists, with top performers (AMZN, META, MSFT, AAPL, GOOGL) each appearing 34 times.
- **Price Category Performance**: Mid-price ($10-$50) stocks showed highest average gains (>10%), while High-price stocks demonstrated more modest but consistent 1-2% gains.
- **Day Patterns**: Monday, Friday, and Tuesday had highest stock appearance frequency, suggesting optimal monitoring days.
- **Performance Correlation**: Frequently appearing stocks (30+ appearances) showed more stable performance compared to less frequent gainers.

Conclusion:

This analysis enables non-financial stakeholders to identify reliable performers, select investments based on risk preference, optimize timing of investment activities, and assess reliability through cross-referencing metrics.The modular data structure allows for future expansion to incorporate additional dimensions, which may help specify what non financial stakeholders are looking for.

Ultimately, this is a classroom exercise on a week's run to capture data. This exercise has been helpful in allowing students to experiment with new tools(for some) and build a pipeline for zero. In our classroom experiment with the fictional stakeholders this project has shown us as data scientists that we are stewards of the data, we can only show what is there in a sense. But it is important to convey that idea to clients in a way that is not negative, but honest. This allows for understanding and transparency in our field. To continue with this project for the stakeholders a bigger data run would be needed. Also it may be important to note the political strife occurring these past few months, which have certainly impacted forecasted stock behavior. This brings up a good point to have a multi sector approach when tackling these problems, as outside factors can bring in "noise" that has to be understood in order to move forward with data analysis.

ERD:

WSJGAINERS_NORM

string	SYMBOL
float	PRICE
float	PRICE_CHANGE
float	PRICE_PERCENT_CHANGE

YGAINERS_NORM

string	SYMBOL
float	PRICE
float	PRICE_CHANGE
float	PRICE_PERCENT_CHANGE

combined into

GAINERS_CONSOLIDATED

string	SYMBOL
float	PRICE
float	PRICE_CHANGE
float	PRICE_PERCENT_CHANGE
date	DATE
string	SOURCE

analyzed for frequency · *analyzed for distribution* · *analyzed by day* · *compared by source* · *analyzed for recurrence*

SYMBOL_FREQUENCY_ANALYSIS

string	SYMBOL
int	APPEARANCE_COUNT
date	FIRST_APPEARANCE
date	LAST_APPEARANCE
float	AVG_PERCENT_CHANGE

PRICE_DISTRIBUTION_ANALYSIS

string	PRICE_RANGE
int	COUNT
float	AVG_PERCENT_CHANGE
date	ANALYSIS_DATE

DAY_OF_WEEK_ANALYSIS

string	DAY_OF_WEEK
int	UNIQUE_SYMBOLS
float	AVG_PERCENT_CHANGE
int	TOTAL_GAINERS

SOURCE_COMPARISON_ANALYSIS

string	SOURCE
int	GAINER_COUNT
int	UNIQUE_SYMBOLS
float	AVG_GAIN
float	MIN_GAIN
float	MAX_GAIN

RECURRING_SYMBOLS_ANALYSIS

string	SYMBOL
int	APPEARANCE_COUNT
float	AVG_PERCENT_CHANGE
date	FIRST_APPEARANCE
date	LAST_APPEARANCE

contributes to · *enhances* · *provides details for*

COMPREHENSIVE_ANALYSIS

string	SYMBOL
int	APPEARANCE_COUNT
float	AVG_PERCENT_CHANGE
string	PRICE_CATEGORY
int	DAYS_BETWEEN_APPEARANCES
int	SOURCES_COUNT
boolean	RECURRING_FLAG
string	DAYS_OF_WEEK