AFM 346 Final Report – Stock Value Predictive Analysis

Group #1: Callum Stevenson, Jake Vanderweyst, Joel Palmer, Nathan Farquharson, Tim Sankey

Introduction

This project aims to develop a predictive model using historical stock price data from S&P 500 companies, which represent the 500 largest firms listed on U.S. stock exchanges, to assist a high-profile client in making informed investment decisions. The objective is to analyze the dataset from multiple angles to identify the most effective predictive techniques, build models, and provide strategic investment recommendations on how to allocate $1 million across six target companies:

1. **Salesforce (CRM)** – Provides customer relationship management technology.
2. **Uber (UBER)** – Provides ride-hailing, food delivery, and freight services.
3. **Chipotle (CMG)** – Provides fast-casual restaurants focused on Mexican food.
4. **Kraft Heinz (KHC)** – Produces and markets wide range of food and beverage products.
5. **Estée Lauder (EL)** – Develops and markets a range of personal care products.
6. **Expedia Group (EXPE)** – Provides booking services for various travel-related needs.

For structural purposes, we will be implementing the CRISP-DM model throughout the project. CRISP-DM is a data science life cycle model that provides guidance on the organization of a data science project. The components of the CRISP-DM model are as follows:

1. **Business Understanding:** Details of the project's business, objectives and requirements
2. **Data Understanding:** Preliminary techniques to familiarize ourselves with the data.
3. **Data Preparation:** Initial activities involving preparing data for analysis.
4. **Modeling:** Selecting and applying various modeling techniques.
5. **Evaluation:** Assessing the models to ensure they meet business objectives.
6. **Deployment:** Practical application of the model to solve the business problem.

By leveraging advanced feature engineering, exploratory data analysis, and various predictive modeling methods, the project aims to deliver actionable insights and maximize the client's potential returns within a specified timeframe. The aim is to predict stock performance for the week of July 22, 2024. We will invest $1 million at the opening price on July 22 and sell at the closing price on July 26 for short-term profit. Our model will predict the value of our investments this week and allocate the portfolio's composition based on the best-performing stocks.

Business Understanding

The business understanding component has been detailed above, providing the foundation for our approach. However, there are some key factors to consider in order to fulfill the objectives and requirements of the task at hand.

Potential factors impacting our investment decision:

- **Market Trends**: Overall market conditions and trends significantly influence stock prices. Understanding the macroeconomic environment is important.
- **Company Performance**: Individual company performance, including financial health, management decisions, and business strategies, directly impacts stock prices.
- **Industry Trends**: Each company operates in a different industry, and sector-specific trends can affect stock performance.
- **Economic Indicators**: Interest rates, inflation, employment rates, and GDP growth are some economic indicators that can influence market sentiment and stock prices.
- **Investor Sentiment**: Market sentiment and investor behaviour, driven by news, social media, and market rumors, can cause stock price volatility.
- **Technical Indicators**: Historical price patterns and technical indicators (e.g., moving averages, RSI) are used to predict future price movements.

Our focus will be on predictive models based on historical data. However, external news events could impact our model's accuracy by introducing factors that the models might struggle to pick up. For instance, Chipotle has an earnings report scheduled for July 24, which could significantly affect its stock price during our investment week. For this reason, it is also important to consider possible alternative events that could impact our investment decision.

Data Preparation and Data Understanding

The dataset consists of historical stock values of S&P 500 companies from the beginning of 2020 through June 27, 2024. It contains 568,390 observations and 8 columns consisting of the variables:

- **Date** – trading date
- **Symbol** – company ticker
- **Adj Close** – adjusted price at market close
- **Close** – price at market close
- **High** – max trading price during day
- **Low** – min trading price during day
- **Open** – price at market open
- **Volume** – volume traded during day

We explored a detailed exploration of the descriptive statistics to gain insights into the distribution and summary metrics of the data. This included examining key statistics such as means, medians, and standard deviations. In addition, we assessed the overall quality of the dataset, identifying and documenting the presence of missing (NA) values. All of the NA values are companies that were added or dropped from the dataset (S&P500) within a span of time that does not align with the first and last day of the time-period that is being analyzed. To maintain the integrity of our final model and avoid potential skewing from imputation, we decided to remove all missing (NA) values from the dataset. Additionally, we cleaned the data by sorting the data frame by date and stock ticker symbol to ensure that all analyses and subsequent modeling are based on a well-organized and accurate dataset.

To enhance the analysis of stock data, we created new variables using various feature engineering techniques. These new variables provide deeper insights into stock price movements and trends, enabling more informed decision-making. Below are the new variables created to enhance our analysis:

- **Date time**: a new datetime-formatted variable derived from the existing 'Date' column to facilitate time-series analysis, recognizing that the data only contains weekdays, meaning that one week is equivalent to five days.
- **Lag**: allow us to incorporate past values of a time series into our analysis, enabling the modeling of temporal dependencies, trends, and patterns.
- **Moving Average**: allow us to smooth out short-term fluctuations in stock prices to identify underlying trends over different time periods, making it easier to analyze the general direction of the market.
- **Moving Standard Deviation:** enable us to measure the volatility of stock prices over different time periods, providing insights into the stability and risk associated with the stock.
- **Return**: allow us to analyze and compare the stock's performance over different time horizons, identify trends, assess volatility, and make informed decisions based on historical price movements and returns.
- **Moving Averages of Volume**: these calculations smooth out short-term fluctuations to identify underlying trends in trading activity over different periods, such as 5-day and 10-day periods.
- **Day of Week, Month, and Quarter**: extracting these attributes captures seasonal and cyclical effects, enabling the analysis of how stock performance varies depending on the time of the week, month, or quarter.

The technical analysis library was used to provide functions for calculating various technical indicators, which are used to analyze stock price movements and trends, the following technical variables, indicators and interaction features were created:

- **Relative Strength Index**: calculated with a 14-day window to measure the speed and change of price movements.
- **Bollinger Bands**: calculated with a 20-day window to provide upper and lower bands based on standard deviations from the moving average.
- **Moving Average Convergence Divergence**: calculated using a slow window of 26 days and a fast window of 12 days to identify changes in the strength, direction, momentum, and duration of a trend.
- **Exponential Moving Averages**: calculated with 12-day and 26-day windows to give more weight to recent prices, providing a smooth trend.

- **Stochastic Oscillator**: calculated with a 14-day window to compare a particular closing price to a range of its prices over a certain period.
- **Average Directional Index**: calculated with a 14-day window to quantify the strength of a trend.
- **Commodity Channel Index**: calculated with a 20-day window to identify cyclical trends in a security's price.
- **Volume Price Trend**: combines price and volume to identify the direction of price movements.
- **On-Balance Volume**: uses volume flow to predict changes in stock price.
- **Accumulation/Distribution Index**: uses the relationship between the stock's price and volume to identify divergences between price and volume flow.

We revisited the analysis of missing (NA) values to check for any errors with the newly added variables. The presence of NA values in the lag and moving average variables was expected as these are due to the timing and insufficient data available to compute lagged values for these observations.

We created subsets of the data set to facilitate more focused analysis which included:

- **Target Companies Subsets**: we created a separate subset of the main data frame for the group of target companies. This allowed us to focus down and explore just the targeted companies we are potentially investing in.

- **Individual Company Subsets**: we created separate subsets of the main data frame for each individual target company to help with targeted analysis and modeling for each individual company.

- **Comparable Company Subsets**: we created subsets of the main data frame for companies that are comparable to the target companies. This allows for benchmarking and comparative analysis by grouping each target company with its peers in the industry. By analyzing these subsets, we can gain insights into the relative performance and trends of the target companies against their respective competitive sets.

For the initial pattern exploration to investigate the stock price performance of the target companies, we developed various visualizations and explored multiple analytical sections:

- The first section focuses on comparing overall stock price trends and performance over time. A time series plot showing the stock price changes for each company over the specified period, allowing for a comparative analysis of their trends and overall performance (Exhibit 1). A plot of the monthly dollar amount change in selected stock prices, highlighting the absolute fluctuations and identifying periods of significant price movements (Exhibit 2). A plot of the percentage change in selected stock prices on a monthly basis, enabling a relative comparison of the volatility and growth rates across the different companies (Exhibit 3).

- The second section focuses on analyzing the stock performance during the week we intend to invest, based on data from previous years. These visualizations help identify patterns and trends specific to that week, providing insights for investment decisions (Exhibits 4-9). Note that the week of July 26 is highlighted in red on the graphs.
- The third section analyzes each selected stock by displaying the actual adjusted closing price over time, the 20-day rolling mean (smoothed trend), and the 20-day rolling standard deviation (volatility). These graphs provide important information for short-term investors, showing trends and volatility for more informed decision-making (Exhibits 10-15).

Modelling

Note that all models' results can be seen in the next section

Model #1

Our first model focuses on predicting the stock value through historical data points using linear regression. Linear regression was chosen as we tested the MSE value of all the models that gone over in class and linear regression was consistently the lowest score. Testing the R-squared value for each variable we found that lag1, MA_5, MA_10, EMA_12, and lag5 were consistently strong at covering the variability of the price so these 5 variables will be used in the linear regression. We observed that each stock's model had an R-squared >0.98 and the graph showed a close match between actual and predicted prices. Consistently, the cross-validation score was >0.96 ensuing robustness and accuracy of the model.

Using the model, we predicted out to the last known stock price on 2024-07-19, this allowed us to sensitivity test our model with real data to ensure the prices were realistic in a real-world setting. The prices were predicted using a rolling window approach which considers each day of predicted values for the next day. We also added a random shock based on the stock history to try and create more realistic unpredictability in fluctuations that we observe in the open market. From here we predict the price on 2024-07-26 and calculate a return for our week

Model #2

The second model was used to forecast stock prices 21 trading days ahead, using a forward lag variable to capture the gap between the end of the historical data and the sale date. Our approach focused on selecting the most influential variables and the best-performing model to achieve the greatest accuracy. We predicted prices based on the latest five days of available data to create a detailed forecast for the week of July 22, 2024, to July 26, 2024. The forecasted prices, based on last month’s data, offer valuable insights into expected market trends and will help us determine an appropriate price to assist in the allocation of funds in the portfolio.

To create this model, we created a forward looking variable to align our predictions with the target sale date. A variable selection function was defined to identify the significant predictors of future stock prices for each company. With these variables selected, we iterated through various models, evaluating their performance against each other to identify the top performers. The best-performing models, often linear regression, were trained on the selected variables and used to predict next week’s prices based on the corresponding historical data available.

Model #3

The third model employs a range of technical indicators that capture momentum, volatility, and trend strength. These indicators help identify overbought and oversold conditions, market volatility, cyclical patterns, and buying or selling pressures. The model uses variables such as Stochastic Oscillator, Relative Strength Index, Moving Average Convergence, Average True Range, Fourier Transform, and others. The prediction period spans the next 20 trading days from July 19, 2024, with a validation period of the last 60 trading days.

The prediction model employs Gradient Boosting Regressor and is optimized through GridSearchCV to identify the best hyperparameters. The model was trained on historical data stock data and validated using cross-validation to ensure the robustness and accuracy of the model. The model runs for 20 iterations, each representing a trading day, predicting the change in stock price based on the last known data. The predicted price is added to the last know price to get the new price. The new price is used to update the feature set for the next prediction.

Evaluation

In the below table we are showing the output from the models discussed above. For each model, it lists the actual closing price on July 19, 2024, alongside the projected closing prices for July 26, 2024. The "Return (%)" column calculates the percentage change between the actual closing prices on July 19 and July 26, indicating the expected return over that period. We used percentage return to standardize them across the stocks regardless of their current price, our goal is to maximize the percentage return of our investments.

Stock Performance Chart

Ticker Symbol	2024-07-19 Actual Closing Price	Model Used	2024-07-26 Projected Closing Price	Return (%)	R Squared	MSE
CRM	$247.63	1	$247.71	+0.03%	0.9881	23.09
		2	$224.08	-9.51%	0.5269	517.44

		3				
		Average		**-4.74%**		
UBER	$67.31	1	$69.73	+3.60%	0.9932	1.35
		2	$68.92	+2.39%	0.2346	55.03
		3	$62.69	-6.86%	0.4466	0.217
		Average		**-0.29%**		
CMG	$53.54	1	$62.68	+17.07%	0.9968	0.38
		2	$65.90	+23.09%	0.3198	15.05
		3	$49.77	-7.04%	0.4747	0.09
		Average		**+11.04%**		
KHC	$33.12	1	$31.78	-4.05%	0.9892	0.20
		2	$33.02	-0.30%	0.4206	9.71
		3				
		Average		**-2.18%**		
EL	$99.18	1	$103.67	+4.53%	0.9949	20.00
		2	$105.77	+6.64%	0.2918	668.32
		3				
		Average		**5.59%**		
EXPE	$135.88	1	$123.22	+9.32%	0.9901	12.00
		2	$86.10	-36.64%	NA	603.81
		3				
		Average		**-13.66%**		

Deployment

To maximize profit, we will allocate $1 million across six target companies based on the predictions from our best-performing models, evaluated by percentage return and accuracy. These models, refined through advanced feature engineering and analysis of key factors such as market trends, company performance, and economic indicators, will guide our investment decisions. By investing at the opening price on July 22 and selling at the closing price on July 26, we aim to capitalize on short-term gains, ensuring the portfolio is allocated to the stocks with the highest projected returns.

Salesforce – 0% Allocation

The models deployed suggested a potential loss for Salesforce as the company may be overvalued, making it an unfavourable investment. The notion of Salesforce being overvalued was echoed through external research, including news articles and discussion boards. Additionally, the models showed high error margins which were evaluated in terms of price to compare between stocks. Salesforce's mean squared error (MSE) in terms of price was slightly above average, suggesting an increased investment risk. Therefore, we allocated 0% to Salesforce.

Uber – 0% Allocation

Our models, despite demonstrating acceptable accuracy, predict that Uber will incur a loss. This forecast aligns with historical performance data that shows Uber consistently delivering negative returns during the week of July 26 over the past four years. The most accurate model (model #3) in terms of MSE indicates a loss of nearly 7%, which casts doubt on potential returns. Given the historical and predictive evidence pointing towards negative performance, we are allocating 0% of the portfolio to Uber.

Chipotle – 66.4% Allocation

The predictive models for Chipotle project a significant potential gain, positioning it as a highly attractive investment. The model with the greatest R-squared value expects a return, and the other model that suggests a potential gain has a favourable MSE in terms of price. The accuracy of these two models gives us confidence in the potential returns for Chipotle. Chipotle's historical volatility around earnings reports presents a unique opportunity, with their next earnings report being on July 24. Ultimately, we are allocating 66.4% of the portfolio to Chipotle, amounting to $664,000.

Kraft Heinz – 0% Allocation

The projections for Kraft Heinz indicate a minor loss, making it a less appealing investment option. While the stock exhibits minimal variability, suggesting limited potential for both gains and losses, the predicted loss outweighs the benefits of stability. Our models show strong predictive performance with low MSE values, reflecting the stock's consistent and predictable performance. To focus on investments with more promising returns, we have chosen to allocate 0% of the portfolio to Kraft Heinz.

Estée Lauder – 33.6% Allocation

The models for Estée Lauder predicted an average gain of 5.59%, supported by strong R-squared and MSE values, which indicate reliable and accurate predictions. This positive outlook is reinforced by both models, with strong accuracy scores, giving confidence in Estée Lauder to generate positive returns. Given these factors, Estée Lauder emerges as a favourable investment choice. We have decided to allocate 33.6% of the portfolio to Estée Lauder, translating to an investment of $336,000.

Expedia Group – 0% Allocation

The prediction models for Expedia Group indicate that the stock is likely to incur a loss, with very poor predictive performance demonstrated by the highest MSE scores and R-squared percentages. These unfavourable conditions suggest that investing in Expedia Group could lead to undesirable outcomes. Given the models' poor performance and negative return projections, we have decided to allocate 0% of the portfolio to Expedia Group.

Calculation of Portfolio Allocation

CRM	UBER	CMG	KHC	EL	EXPE
0%	0%	66.4%	0%	33.6%	0%

Estée Lauder

- Positive Return Projection: +5.59%
- 5.59% of the total combined positive return (16.63%) = 33.6%
- Investment Amount: 33.6% of $1 million = $336,000

Chipotle

- Positive Return Projection: +11.04%
- 11.04% of the total combined positive return (16.63%) = 66.4%
- Investment Amount: 66.4% of $1 million = $664,000

Final Strategic Investment Recommendation

In conclusion, our strategic investment plan allocates the full $1 million across Chipotle and Estée Lauder, based on their strong potential for short-term gains as indicated by our predictive models. Specifically, we are investing 66.4% of the portfolio, or $664,000, in Chipotle, capitalizing on its substantial positive return projection and historical volatility around its upcoming earnings report. Additionally, 33.6% of the portfolio, or $336,000, is allocated to Estée Lauder, reflecting confidence in its projected average gain of 5.59% and the reliability of our models. By excluding Salesforce, Uber, Kraft Heinz, and Expedia Group due to their unfavourable projections and high risk, we are focusing our resources on investments with the highest potential returns. This allocation strategy aims to optimize our portfolio's performance, maximizing profit while effectively managing risk over the short-term investment horizon.

Appendix

Exhibit 1

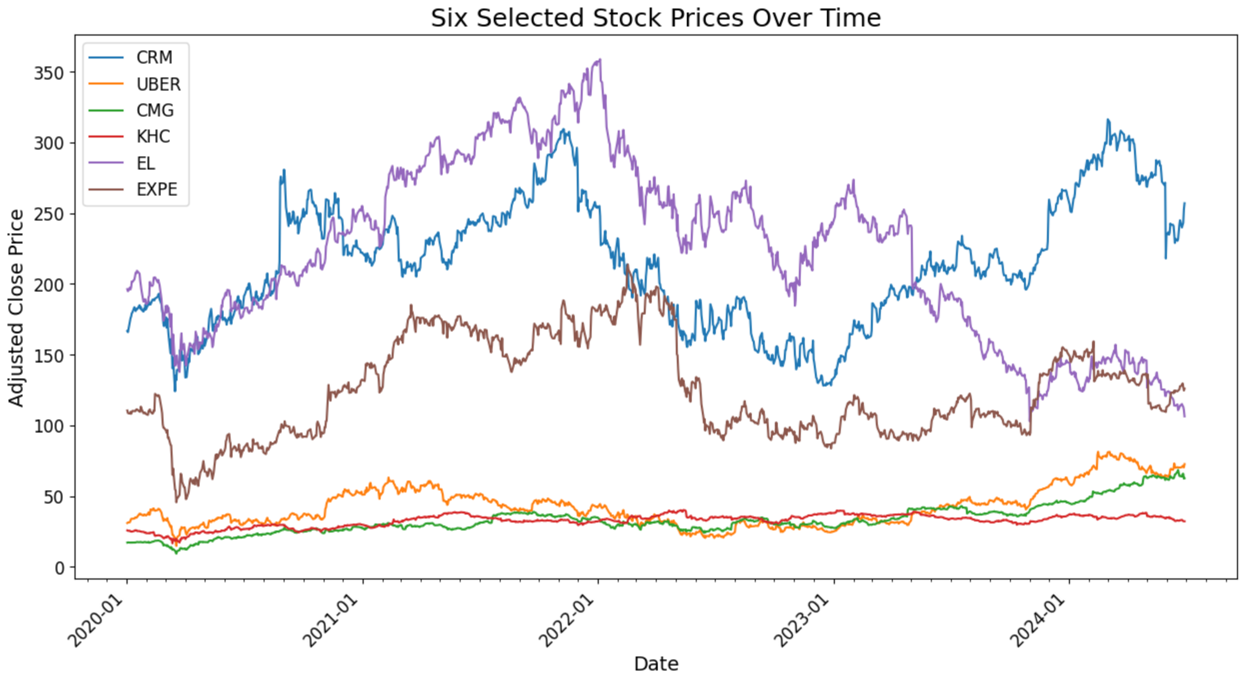


Exhibit 2

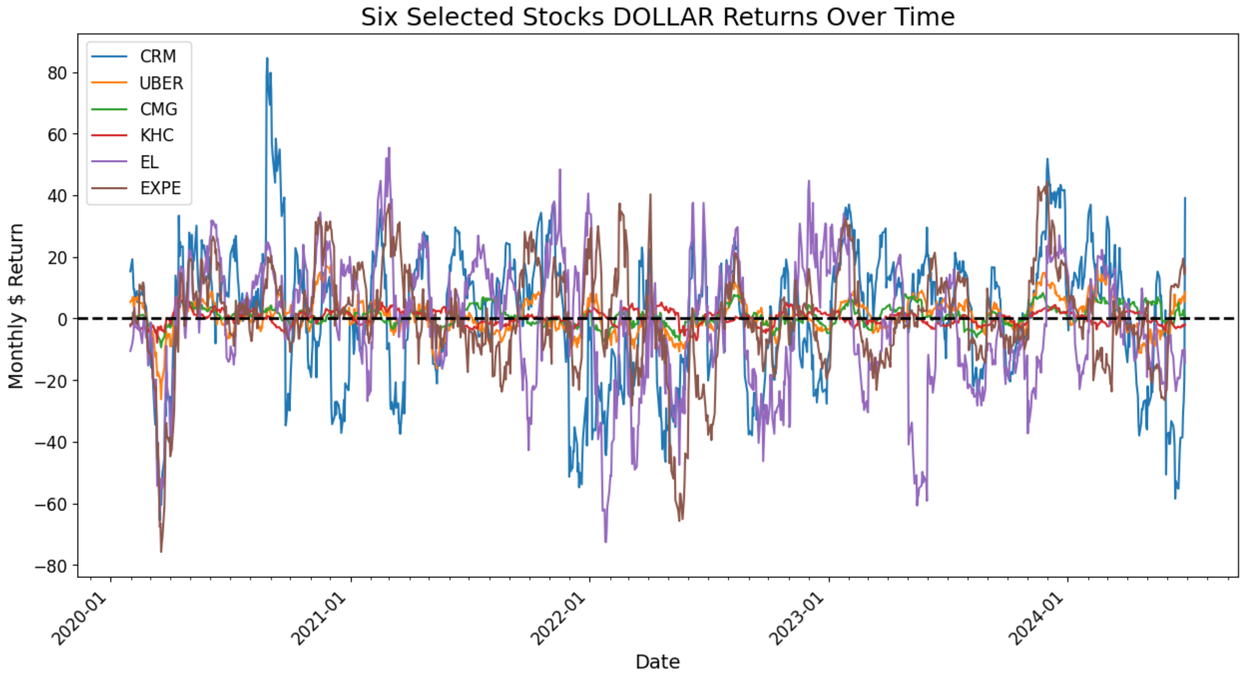


Exhibit 3

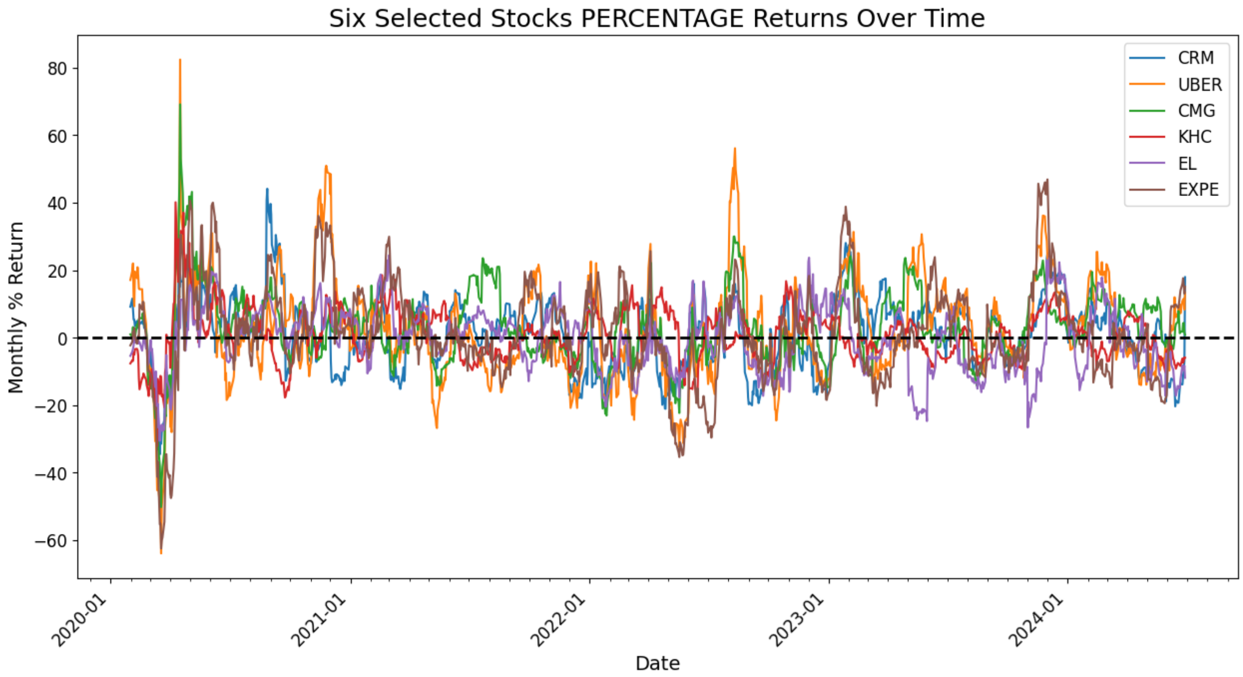


Exhibit 4

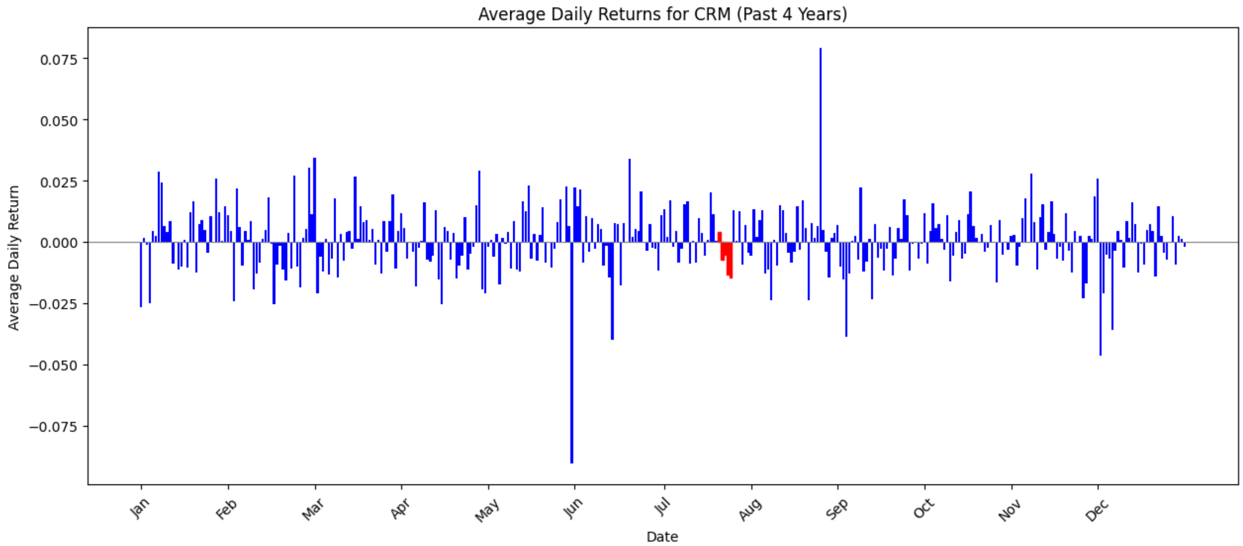


Exhibit 5

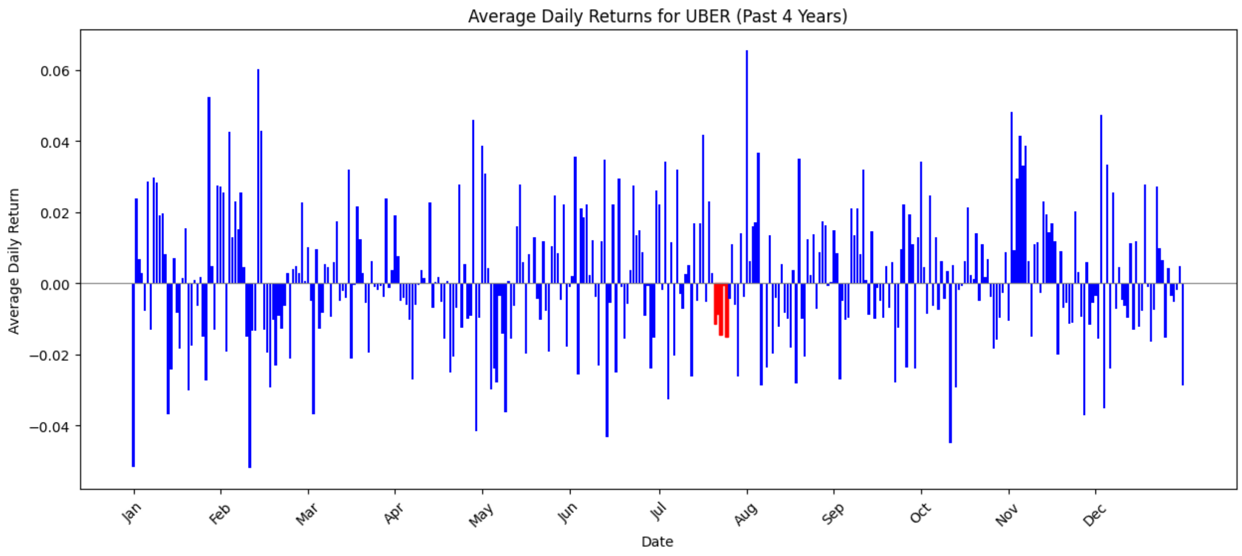


Exhibit 6

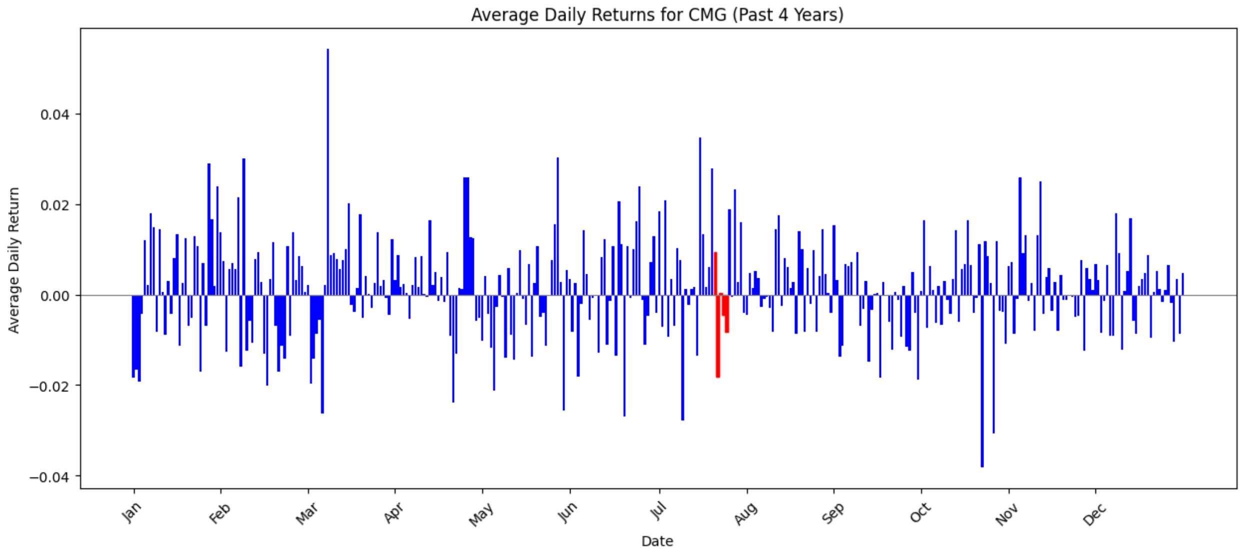


Exhibit 7

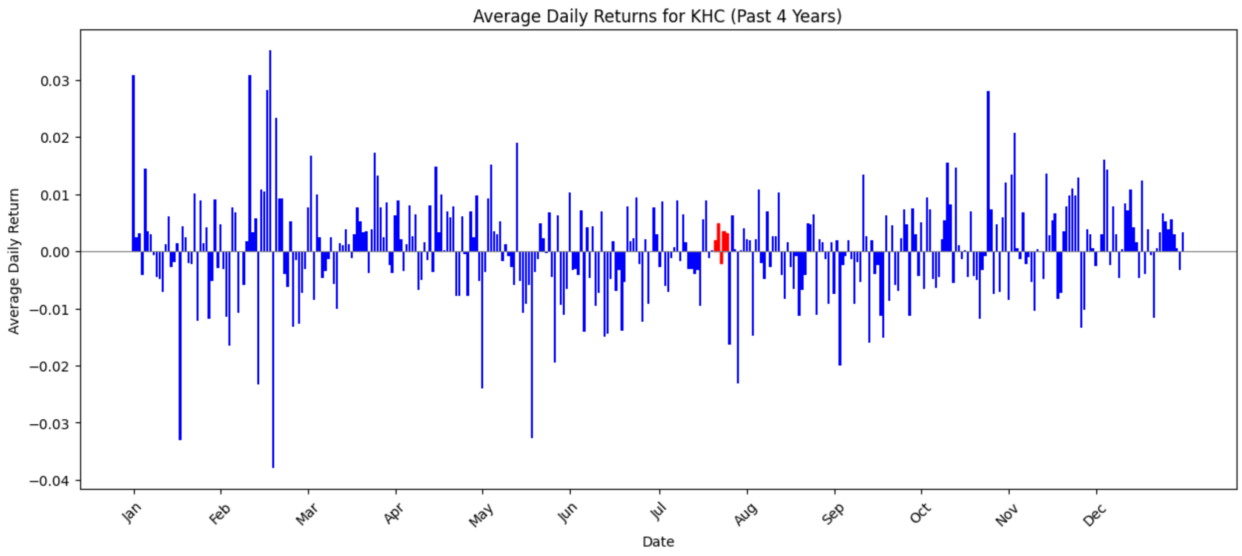


Exhibit 8

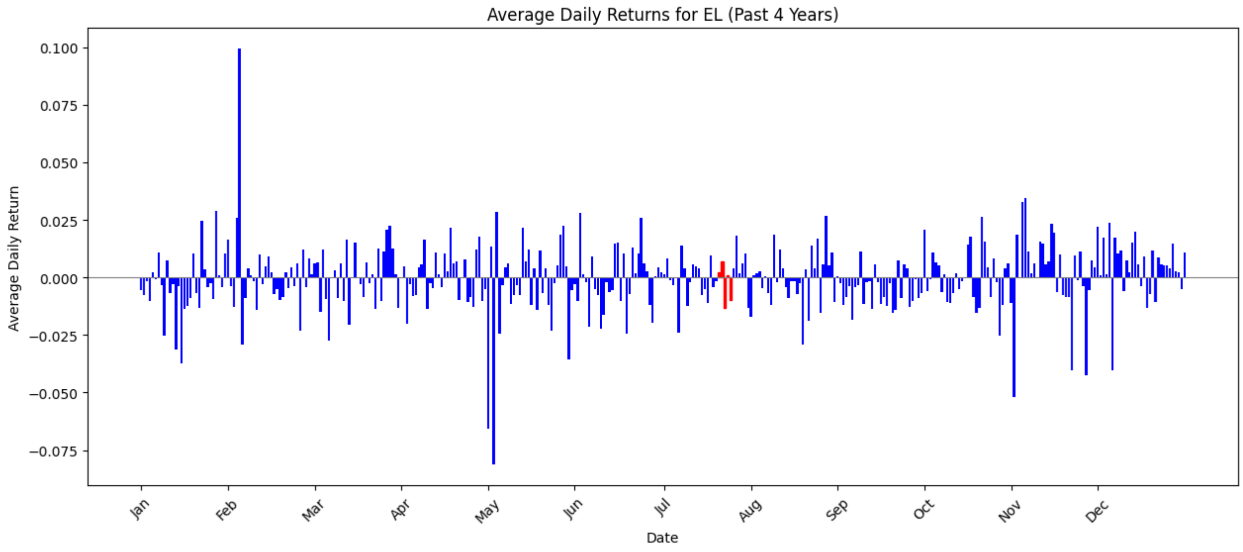


Exhibit 9

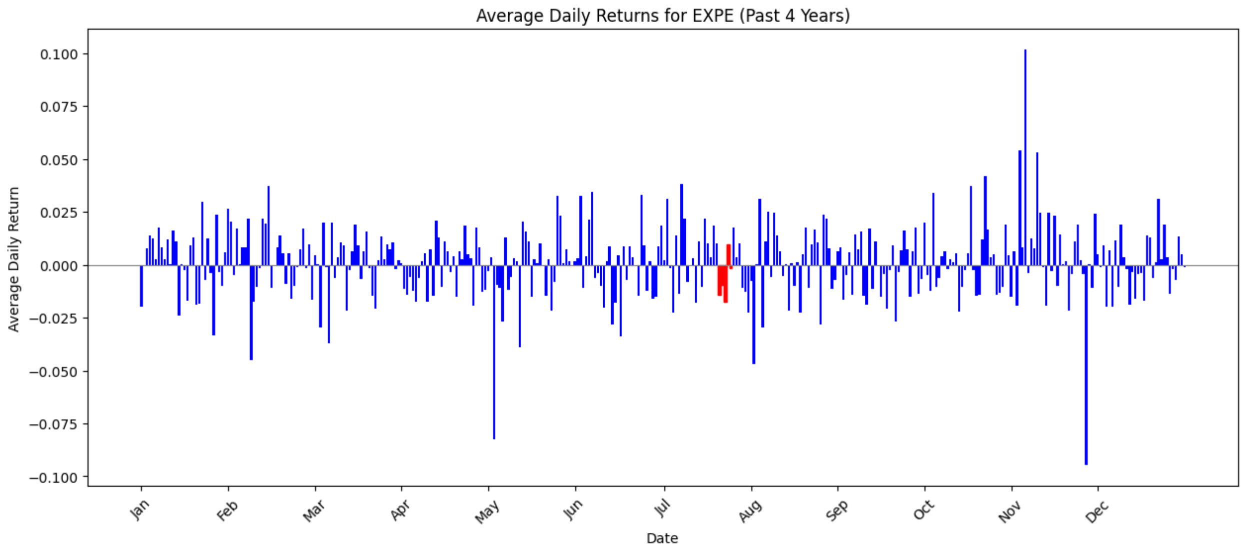


Exhibit 10

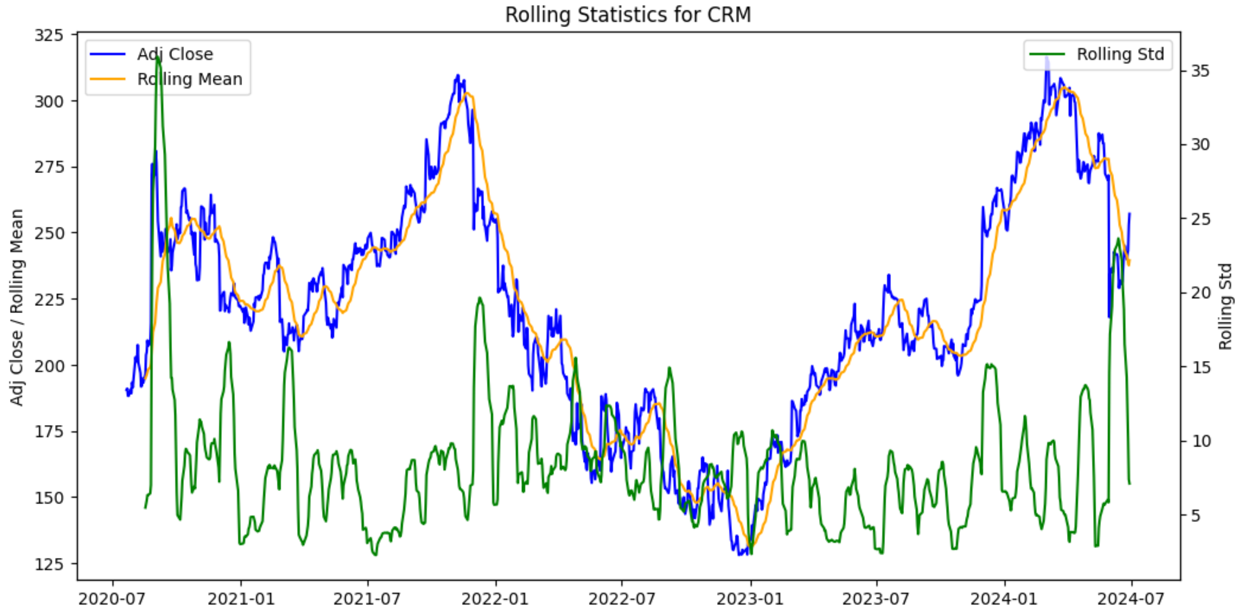


Exhibit 11

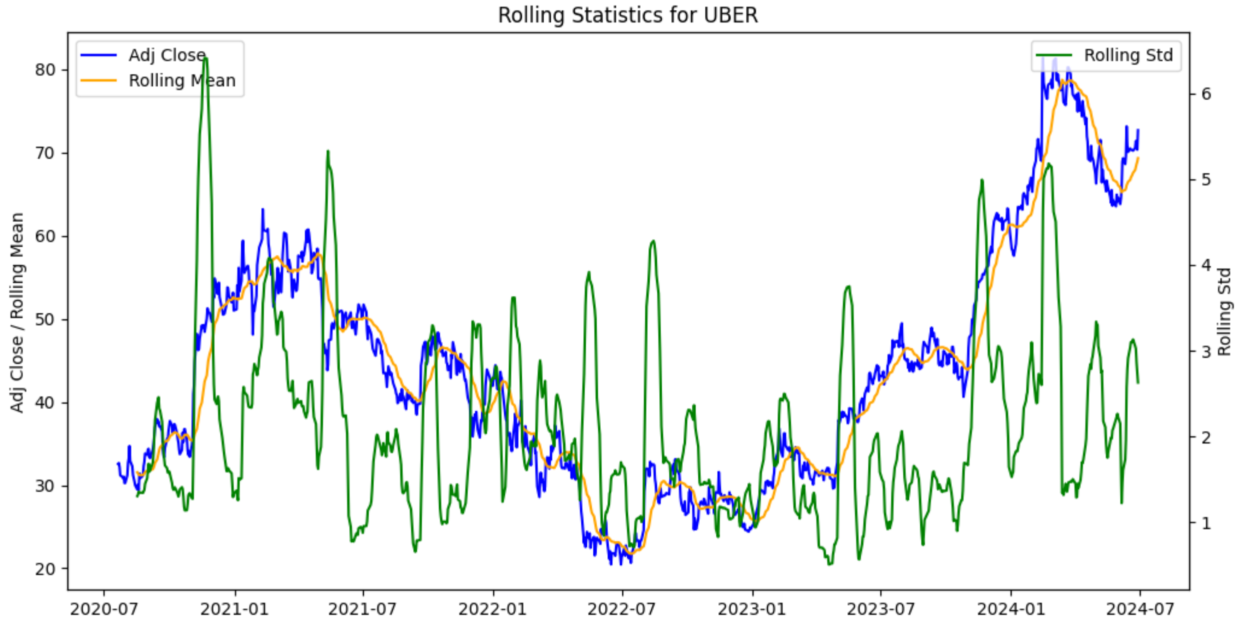


Exhibit 12

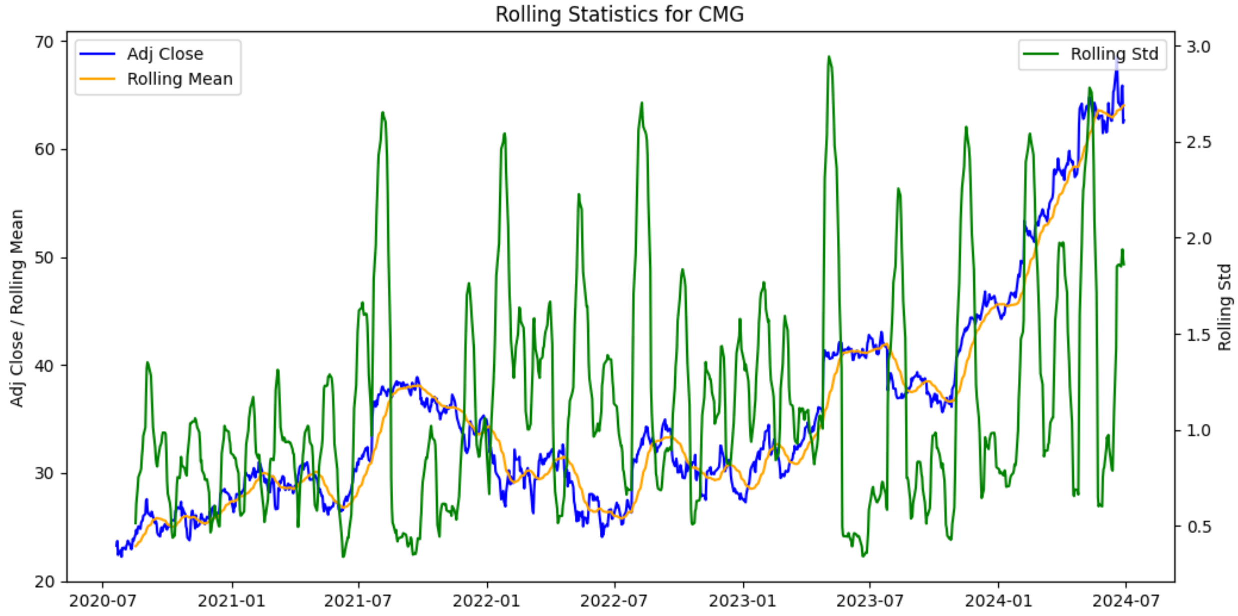


Exhibit 13

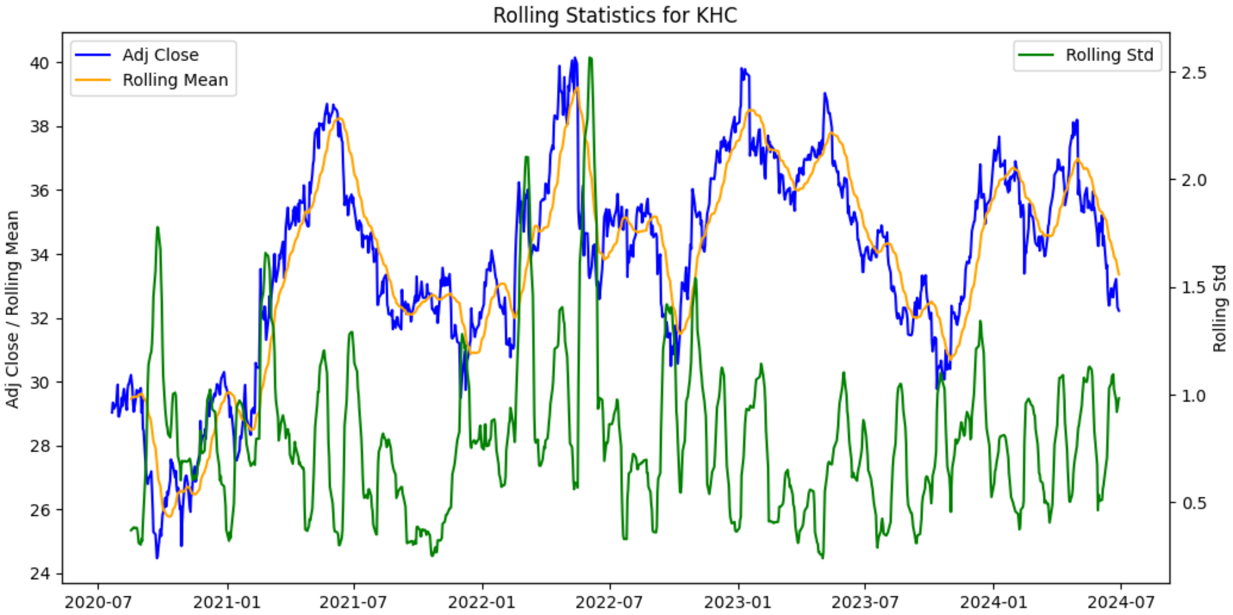


Exhibit 14

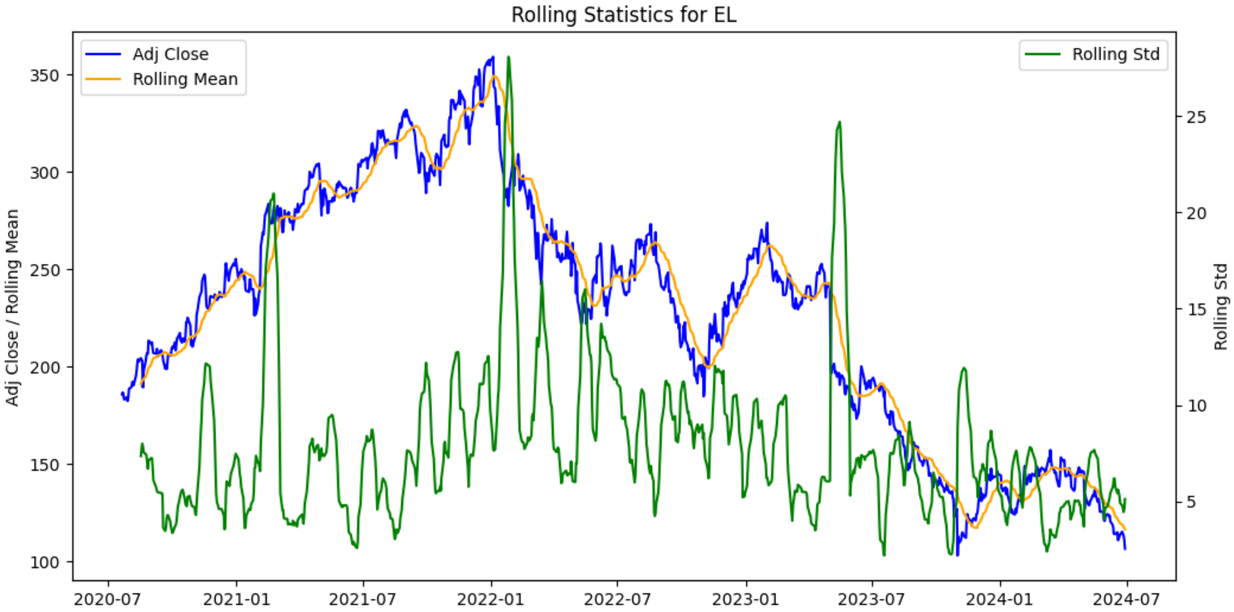
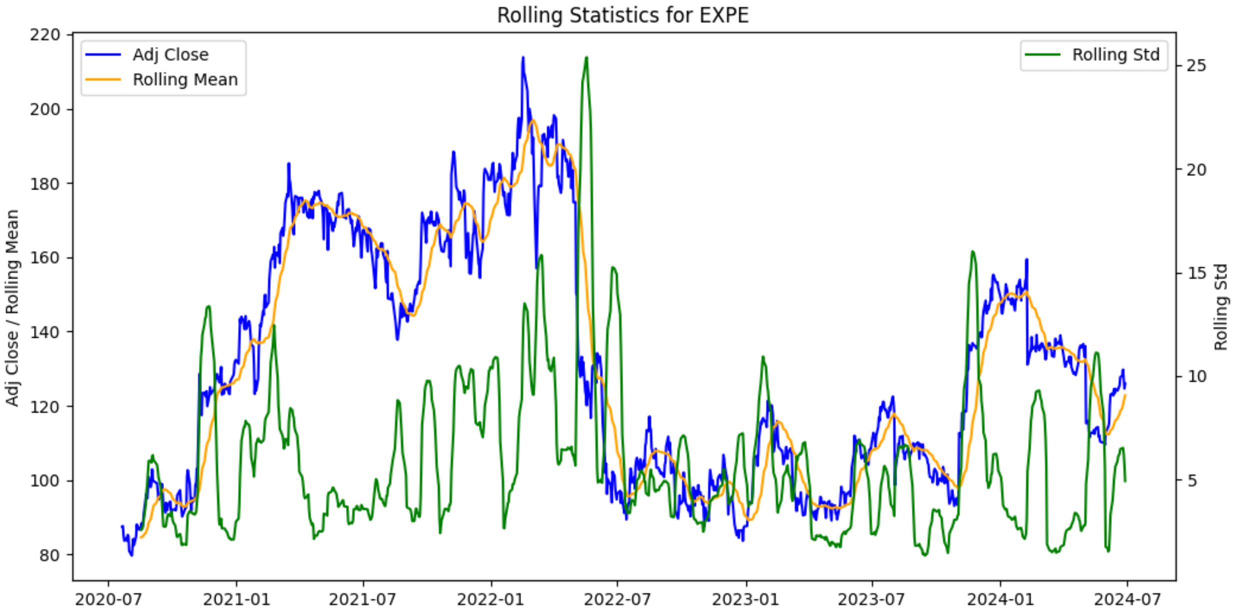


Exhibit 15


Rolling Statistics for EXPE
Adj Close
Rolling Mean
Rolling Std
Adj Close / Rolling Mean
Rolling Std
220
200
180
160
140
120
100
80
25
20
15
10
5
2020-07
2021-01
2021-07
2022-01
2022-07
2023-01
2023-07
2024-01
2024-07