Modeling Stocks By Sector

Joseph Hudson, Joseph Kaminetz, Will Novak, Will Brannock, Aiden Rocha, Garret Knapp, Sam Kunitz-Levy

Missing Data Analysis (Question 1)

NOTE: All Charts and the code to produce them are also viewable in the stocks_missing_analysis.ipynb file

The data was uploaded to Kaggle by Larxel. According to the author, the stock data is sourced from FRED (Federal Reserve Data) and yfinance (Yahoo Finance). The data includes stock price data for companies included in the S&P 500 index. The index is a composite of the 500 largest American companies by market cap. The data spans from 2010 to 2024.

There are three files in the data:

- sp500_stocks.csv - this contains all of the daily stock data for the individual companies in the S&P 500 index. The daily data for each stock includes adjusted close, close price, high, low, open, and volume. The days included for each year are the trading days, which excludes weekends, holidays, etc.
- sp500_companies.csv - this is the metadata for each company. There is metadata like sector, which we use to segment the stocks for the markov chain, as well as market cap, full company name, ebitda, geographic information, and weight in the S&P 500.
- sp500_index - this is the daily S&P 500 values. We assume this is close. From our research, the composite is rebalanced approximately quarterly, according to the market cap of the largest 500 companies.

To investigate the missing data, we began with an overall analysis of the dataset with pandas's describe function. We saw that there was missing data but it looked like the data was fully missing for companies with any missing data. To look into this, we examined the data by year to see if there was a temporal trend in missing (i.e. older data tended to be more missing). However, this was disproven when we found that the number of missing rows was approximately consistent across years.

Next, we checked missing data by company. We started to notice that many companies with missing data had the exact same number of missing observations: 3,768. We realized that it meant they were completely missing. Furthermore, we recalled from the class examples that there were 250 trading days in a year, so we divided by 15 and found that it was an exact match. Thus, we concluded that many of the companies with missing data had completely missing data.

Our hope was that we could build a Markov chain by sector, so next we generated a list of companies with all numeric columns missing, broken down by sector. The table had the sector, # of fully missing companies, and the percentage of companies in that sector missing. From this table, we concluded we had enough companies per sector (>5) to run a meaningful analysis. About 65% of each sector has fully missing data.

Next, we determined that the number of companies with partially missing data, an additional 22 companies, was small enough that we could remove them from the dataset.

Lastly, we created a clean dataset with only the 150 companies that have fully complete data. There are at least 5 companies per sector in the data. We will use this data to build the Markov chains by sector. The missing 352 companies represent a significant limitation with the dataset. If the companies with missing data are somehow systematically different than the companies with complete data it could mean that our modeling of the sectors may not be representative of that entire S&P 500 sector.

Phenomenon Overview (Question 2)

NOTE: All Charts and the code to produce them are also viewable in the data_creation.ipynb and ECDF_Random_Sample.ipynb file

We are attempting to predict stock value changes by sector of the economy. With all stocks, it is useful to understand how likely a stock is going to increase or decrease in price after the previous day's close. It is also useful to project how much a stock will go up or down, after deciding whether it will increase or decrease. But while this information can be incredibly useful for stocks individually, how an individual stock of a company performs doesn't tell us as much about the economy as a whole. By looking at how return values change by sector, we can gain more insight about where there might be growth and decline in the market as a whole.

Like we explained earlier, we are attempting to create Markov chains for each sector of stocks in the market. So, our features are the stock sectors with which we have complete information in terms of their closing numbers for the trading days between 01/04/2010 to 12/20/2024. To create this we grouped by the Date And Sector for each observation, that was originally a stock and their adjusted close on a given day. With this group by, we took the sum of adjusted closes for the stocks in that particular sector for that particular day. That can be displayed by the dataframe below:

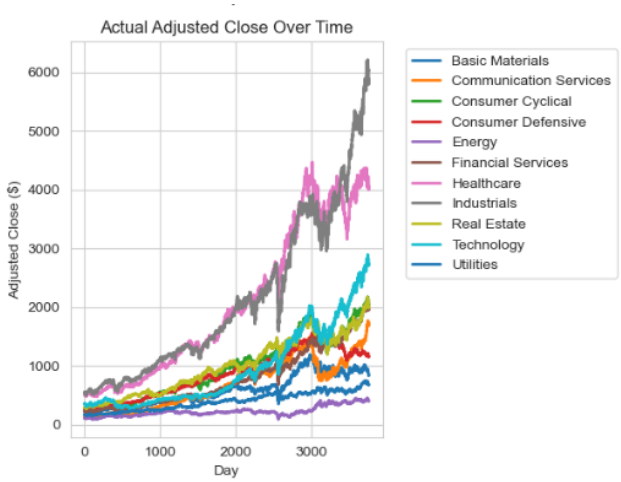
	Date	Sector	Adj Close
0	2010-01-04	Basic Materials	178.189838
1	2010-01-04	Communication Services	128.564411
2	2010-01-04	Consumer Cyclical	241.193282
3	2010-01-04	Consumer Defensive	256.358244
4	2010-01-04	Energy	104.954098
...	...	...	...
41443	2024-12-20	Healthcare	4061.319965
41444	2024-12-20	Industrials	5892.510040
41445	2024-12-20	Real Estate	2011.869986
41446	2024-12-20	Technology	2744.200005
41447	2024-12-20	Utilities	682.539993

From there, we calculated a percentage change for each day that the sector stocks were being traded, that is the percentage change of the stock sectors' closing price compared to the previous day. We also have a column for each day for each sector that is 1 or 0, where a 1 means the sector's closing price went up from the previous close and 0 if the sector's closing price is lower than the previous close. The first day has NaN values because it lacks a prior day. That can be seen below:

	Date	Sector	Adj Close	pctChange	up_ind
0	2010-01-04	Basic Materials	178.189838	NaN	0
1	2010-01-04	Communication Services	128.564411	NaN	0
2	2010-01-04	Consumer Cyclical	241.193282	NaN	0
3	2010-01-04	Consumer Defensive	256.358244	NaN	0
4	2010-01-04	Energy	104.954098	NaN	0
...	...	...	...	...	...
41443	2024-12-20	Healthcare	4061.319965	0.015142	1
41444	2024-12-20	Industrials	5892.510040	0.014285	1
41445	2024-12-20	Real Estate	2011.869986	0.012776	1
41446	2024-12-20	Technology	2744.200005	0.013798	1
41447	2024-12-20	Utilities	682.539993	0.015080	1

After we have computed those percentage changes, we can now look at how our sectors change as a whole from day to day. With this information, we now have the data needed to create Markov chains and KDE/ECDF plots for each sector of stocks in our data.

One key characteristic of this data is that while the sectors may have varying volatility, the overall trend of each sector in the S&P 500 is positive over time. This has significant impacts on our models as detailed in the following sections. This characteristic is evident in the chart below.



Models Overview (Question 3)

For this analysis, we designed two models that allow us to predict the change in the value of a sector of the stock market over time. The first approach was the Markov Chain model, and the second was an Empirical distribution model.

Model 1: Markov Chain

NOTE: All Charts and the code to produce them are also viewable in the Markov_Chain_Work.ipynb file

Creating our transition matrices for each sector was relatively simple. Because we have data for every training day, we can sort each sector by date, and then populate our transition matrices using our column of whether the adjusted closing price increased or decreased that day. Each sector's transition matrix is extremely similar (formatted for docs)

Sector	After an Up (Decrease, Increase)	After a Down (Decrease, Increase)
Basic Materials	[0.4666, 0.5334]	[0.4778, 0.5222]
Communication Services	[0.4555, 0.5445]	[0.4674, 0.5326]

Consumer Cyclical	[0.4431, 0.5569]	[0.4595, 0.5405]
Consumer Defensive	[0.4450, 0.5550]	[0.4689, 0.5311]
Energy	[0.4710, 0.5290]	[0.4815, 0.5185]
Financial Services	[0.4394, 0.5606]	[0.4562, 0.5438]
Healthcare	[0.4302, 0.5698]	[0.4667, 0.5333]
Industrials	[0.4478, 0.5522]	[0.4610, 0.5390]
Real Estate	[0.4631, 0.5369]	[0.4704, 0.5296]
Technology	[0.4515, 0.5485]	[0.4618, 0.5382]
Utilities	[0.4525, 0.5475]	[0.4635, 0.5365]

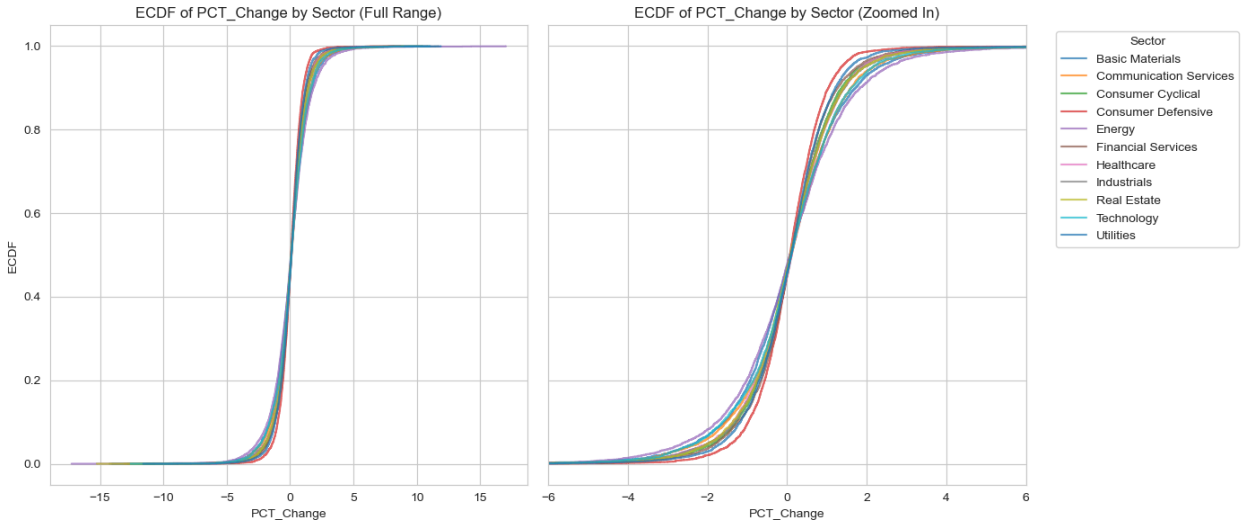
We can take away a few things from just these transition matrices, before we even simulate and generate responses. Firstly, we see that there is little variation in conditional probabilities across sectors. The likelihood for stocks to increase and decrease is almost the same regardless of sector. This might indicate that stock movement, at least just in a pure positive to negative way, is impacted more on the market as a whole as opposed to the industry itself. Another takeaway is that all industries are likely to increase, regardless of whether it increased or decreased the day before. This seems to represent the market's tendency to improve over time. Lastly, consistent across sectors, if a sector has gone up the previous day, the probability that it continues to go up is very slightly higher than on a down-day (and vice versa). These broad takeaways have to be nuanced, as our Markov Chain only indicates whether the stock goes up or down, but doesn't take into account the scale of increases and decreases.

Model 2: Empirical Distribution Model

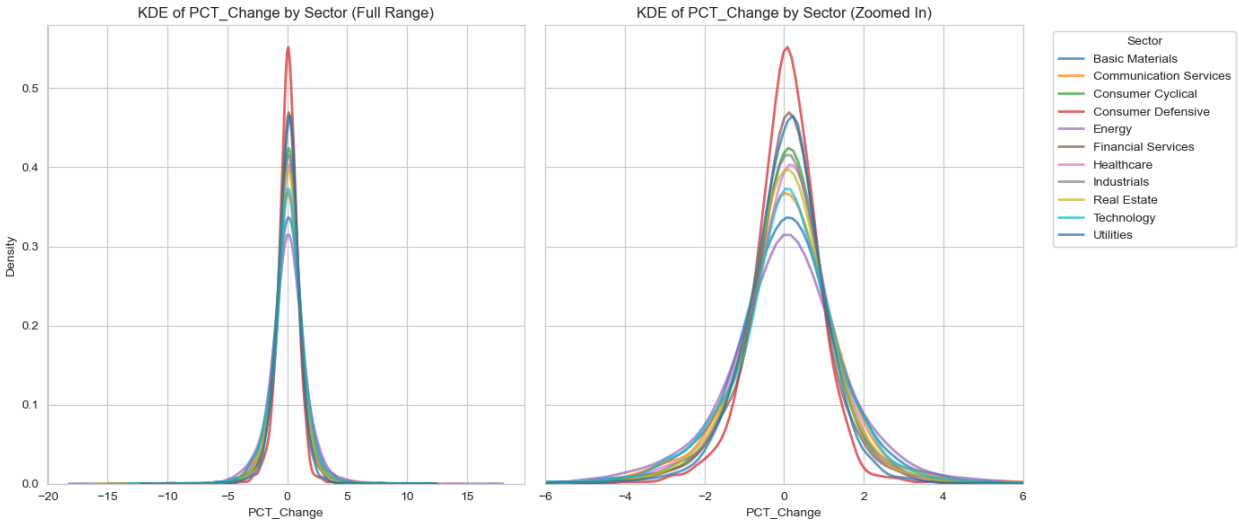
NOTE: All Charts/Code are viewable in ECDF_Random_Sample.ipynb

The goal of this model is to utilize the existing distribution of the daily percentage change in the value of a stock market sector to simulate and predict the value of the sector and gauge the volatility of the sector over time (More in New Data Generation).

The first step was to visualize the Empirical Cumulative Distribution Function and Kernel Density plot to understand the distribution of the daily percentage change across different sectors.



The ECDF plots display the cumulative probability of daily percent changes across different market sectors. Each sector's curve represents how quickly daily returns accumulate toward zero percent change. All sectors exhibit similar S-shaped curves centered around zero, indicating that the majority of daily percent changes are small and symmetrically distributed between gains and losses. The steep central slope reflects that most values cluster tightly near zero, consistent with low day-to-day volatility. Subtle differences in curve steepness and positioning suggest minor variations in volatility across sectors—for example, Energy and Technology appear to have slightly heavier tails, implying a higher frequency of larger daily moves. Overall, the ECDF plots confirm that return distributions across sectors share highly similar shapes, with differences primarily in spread rather than central tendency.



The kernel density plots show the distribution of daily percent changes across different stock market sectors. All sectors display distributions sharply centered around zero, indicating that most daily movements are small and that returns fluctuate closely around no change. The tall, narrow peaks suggest low volatility overall, while the slightly heavier tails visible in the full-range plot point to occasional large daily changes, particularly in sectors like Energy and Technology. Across sectors, the shapes of the distributions are highly similar, implying correlated market movements and consistent behavior across industries. However, subtle differences in peak height and spread suggest that some sectors, such as Consumer Defensive, experience more stability, whereas others, such as Energy, tend to show greater volatility.

Overall, these visualizations provide insight into the underlying data and help infer what to expect from the generated values. The distributions suggest that a sector's value tends to cluster around small positive changes, implying a modest upward trend over time. Accordingly, we expect the simulated data to exhibit a slightly positive slope. It is important to note that the S&P 500 has an all-time positive growth, so this positive slope aligns with real-world behavior.

Additionally, with 3,768 observations per sector, the empirical distributions are sufficiently dense to approximate the true underlying process of daily percent changes. Consequently, random sampling from these distributions provides a statistically sound method for generating simulated sector value changes (View Model 2 Approach and Results).

New Data Generation (Question 4)

Goal

The goal of generating new data with both models is to simulate and compare their predictive performance. Starting from the first day in the dataset, each model is used to create expected daily sector values at two future horizons: **400 days(**Short Term**)** and **~3000 days**(Long Term) from the starting point. For each of these horizons, 5 simulations were run.

After generating these values, the simulations are averaged across runs to obtain a stable expected trajectory for each sector. Finally, the average simulated values are plotted alongside the actual observed sector values to visually and quantitatively assess how accurately each model captures real-world value changes over time.

Model 1 Approach and Results

NOTE: All Charts and the code to produce them are also viewable in the Markov_Chain_Work.ipynb file

To generate new sequences with our Markov chain, we start at the second index of our data for each industry. The value of adj close on the second day for each industry will give us a starting stock price, and whether it grew from the first day for each industry will give us our initial state. This will allow us to test our simulation against what happened historically.

In order to begin our simulation, we first need to answer two questions:

1. How do we determine when the stock should go up or down?
2. How much does the stock price go up or down each day?

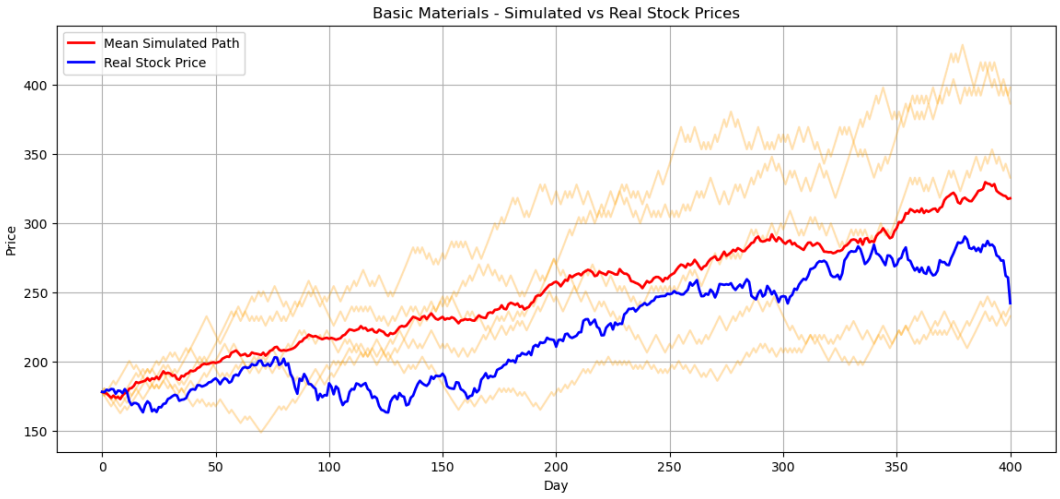
The first question is easy. Because we have our transition matrices, we have values for how likely the stock is to increase or decrease based on the previous state. We first select the correct probabilities for the current state (will the stock go up or down today) based on the previous state (did the stock go up or down yesterday), then we can randomly generate a value from the uniform distribution in between 0 and 1 and if that number is less than our probability value to increase, we set the current state as increase, and otherwise, the stock value will decrease for that day. This works because choosing a value from a uniform distribution between 0 and 1 that is less than a particular value has probability of exactly the value of your threshold. Doing our selection this way ensures that our decision for the stock to increase or decrease is based on the transition probabilities previously calculated.

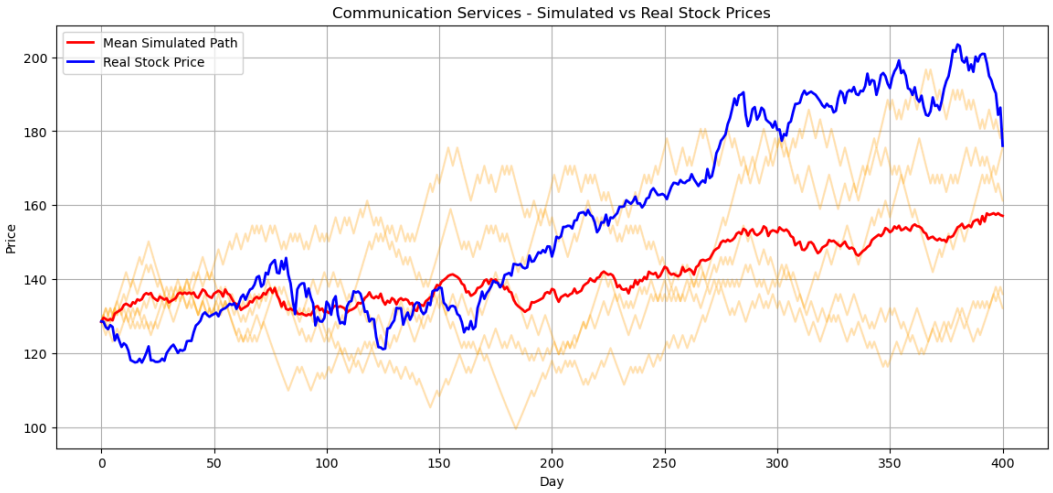
Determining how much the stock price should go up or down each day was extremely difficult, and required some trial and error. At first, we thought our best way was to increase the stock by the average percentage increase found in the data when the stock went up, and decrease the stock by the average percent decrease when the stock went down. This makes logical sense, however, it resulted in our stock prices all simulating to zero in the long run. This is because of the principle of geometric decay. Geometric decay occurs when outcomes are multiplied over time, so random losses outweigh equivalent gains due to compounding. Even if the average return is positive, variability reduces the geometric average growth rate, causing values to drift downward. Instead, we took the standard deviation of the percent changes, which gives us an average daily volatility of the industry. If we exponentiate the standard deviation, we get our up factor, and exponentiate the standard deviation *-1, we can get a down factor that lets us model realistic growth and volatility without experiencing exponential decay, as growth would be additive in log space as opposed to multiplicative.

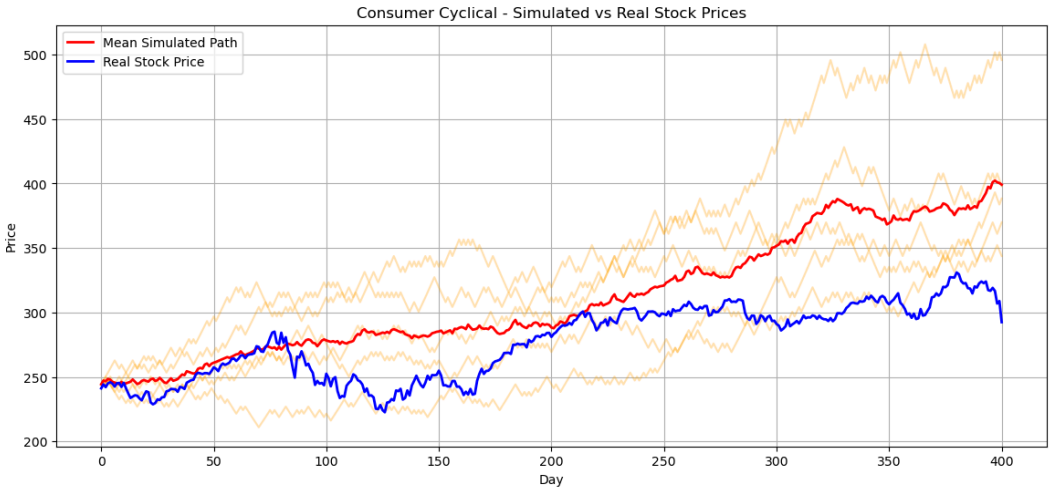
Now that we have our up factor, down factor, and a decision boundary to choose which one should be applied to our stock price, we completed 5 simulations each with 400 days of stock simulation,

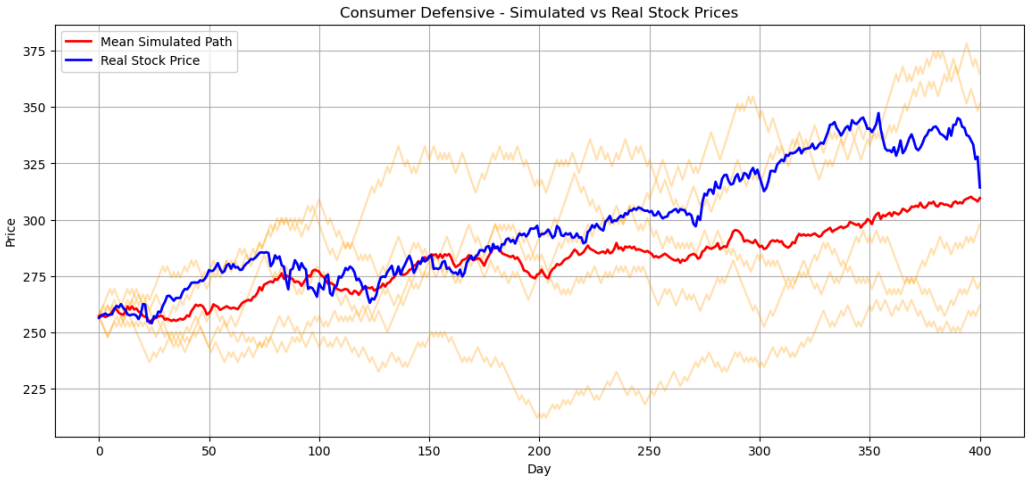
averaged their results to get a short term model, and did the same with 3000 days of stock simulation for a long term model. Those results are compared to the actual stock prices below:

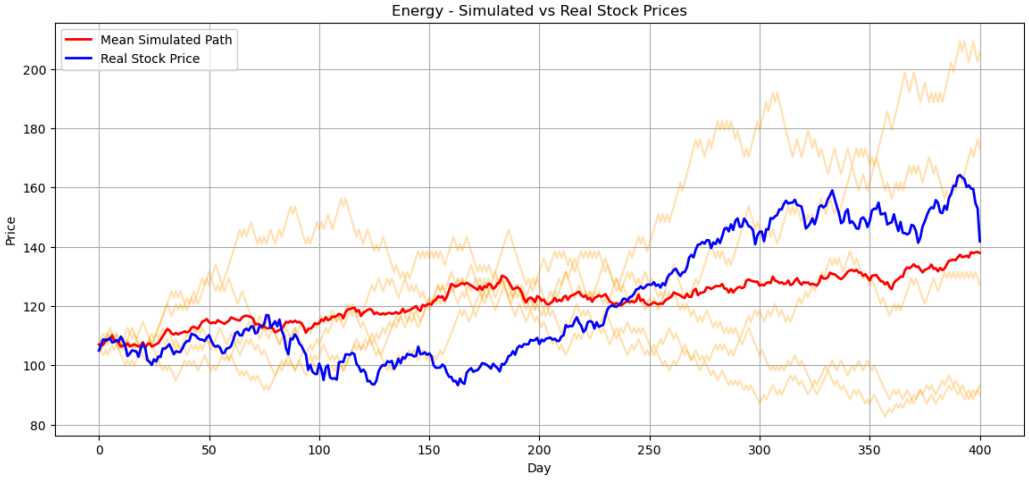
Short Term Model

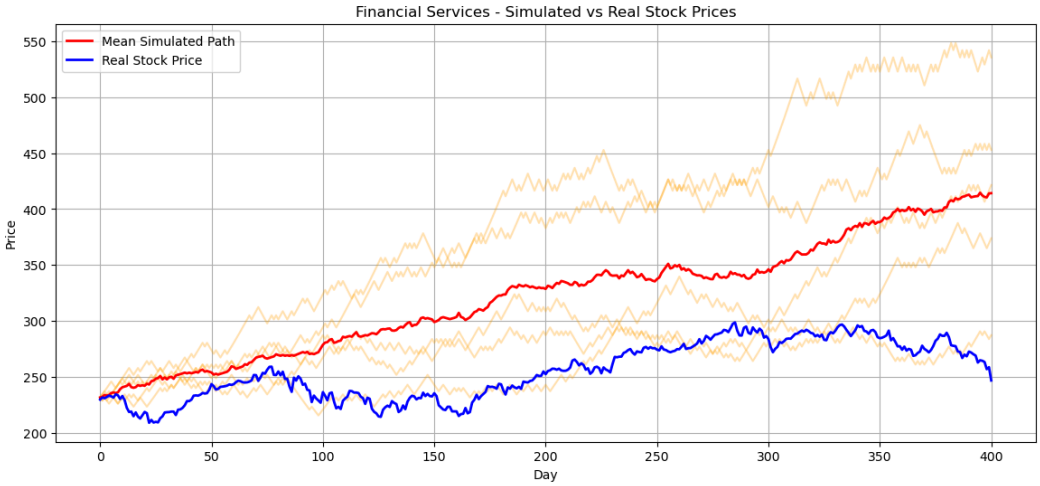


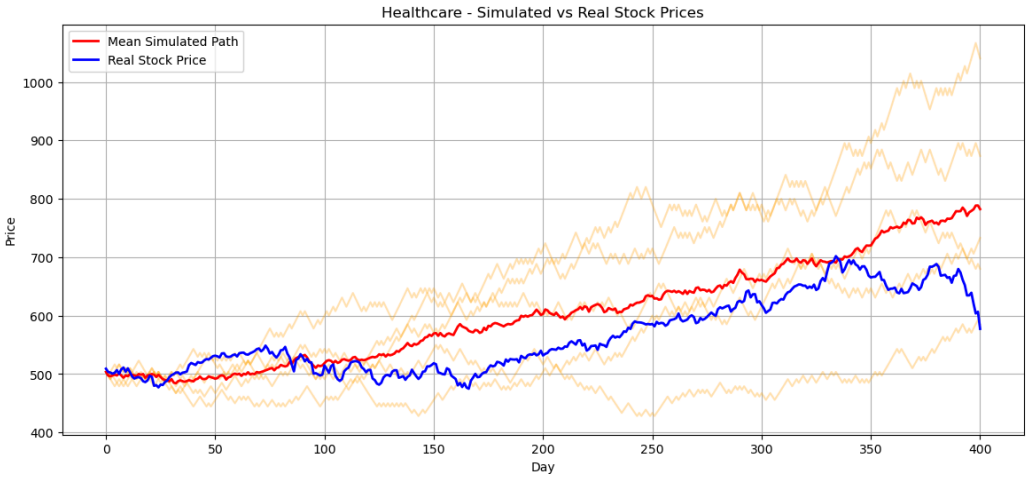


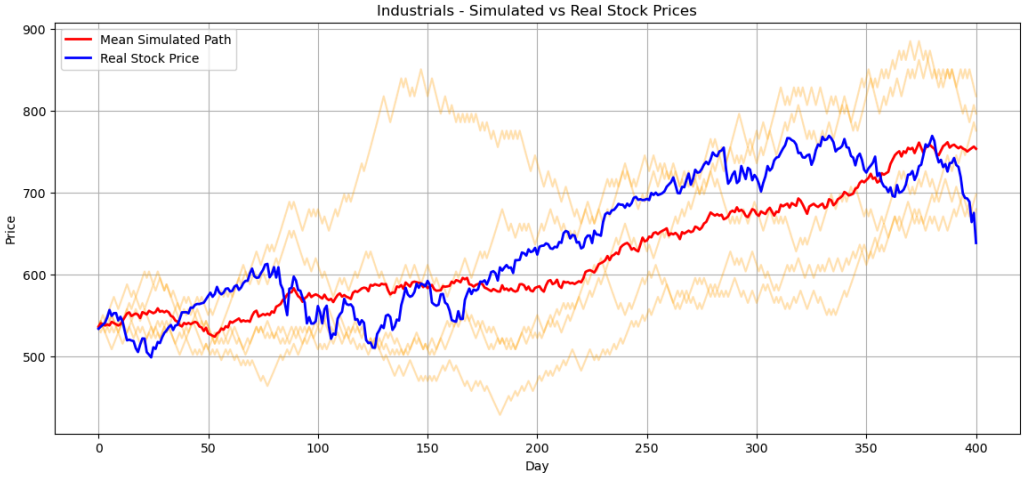


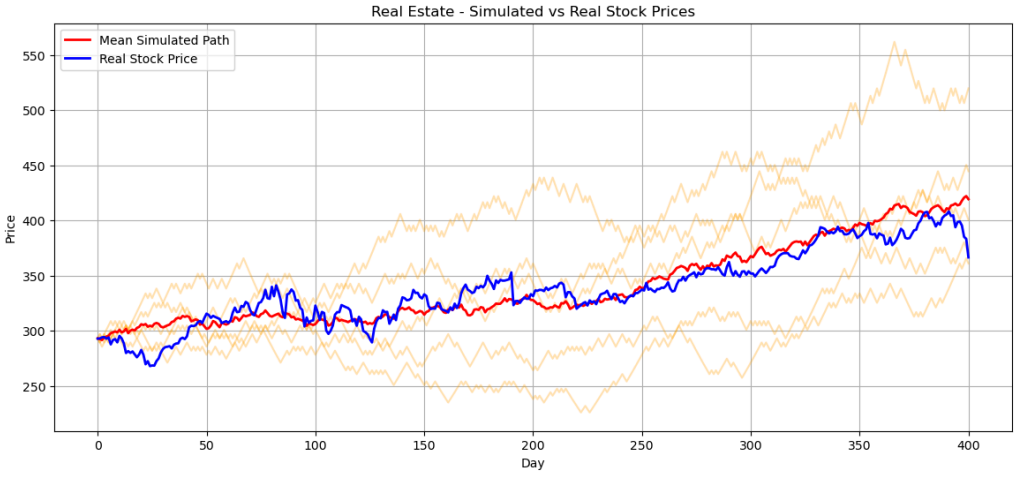

Consumer Defensive - Simulated vs Real Stock Prices
Mean Simulated Path
Real Stock Price
Price
Day

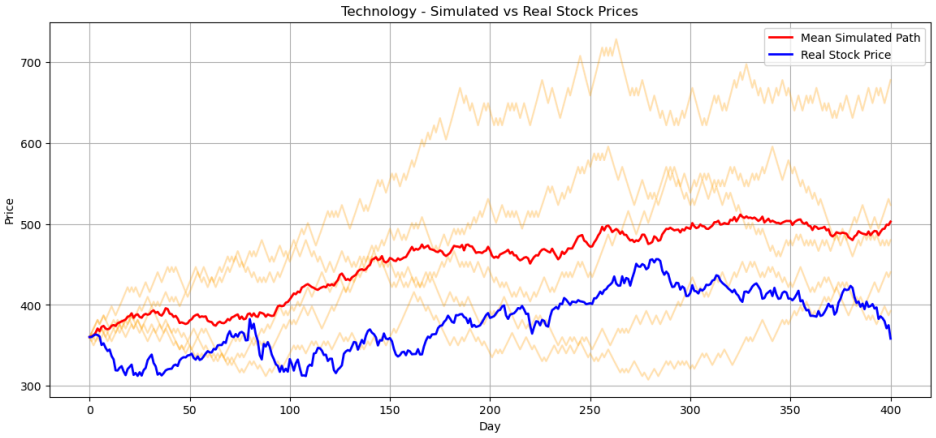

Energy - Simulated vs Real Stock Prices
Mean Simulated Path
Real Stock Price
Price
Day

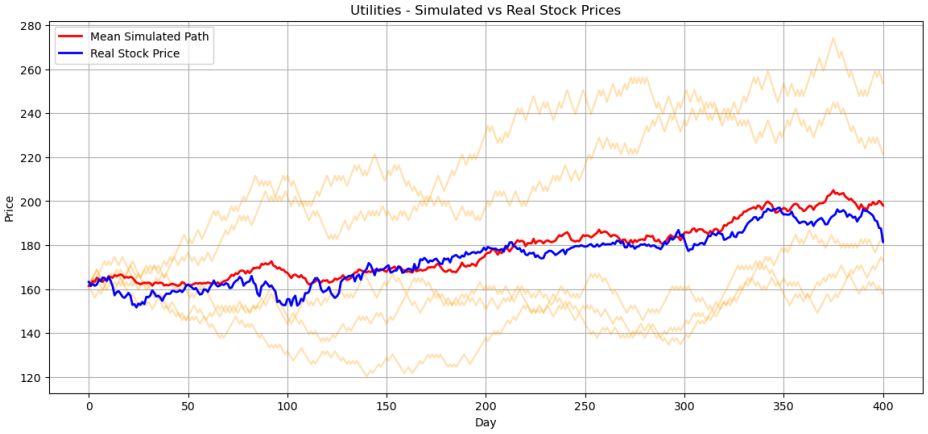

Financial Services - Simulated vs Real Stock Prices
Mean Simulated Path
Real Stock Price
Price
Day


Healthcare - Simulated vs Real Stock Prices
Mean Simulated Path
Real Stock Price
Price
Day
400
500
600
700
800
900
1000
0
50
100
150
200
250
300
350
400

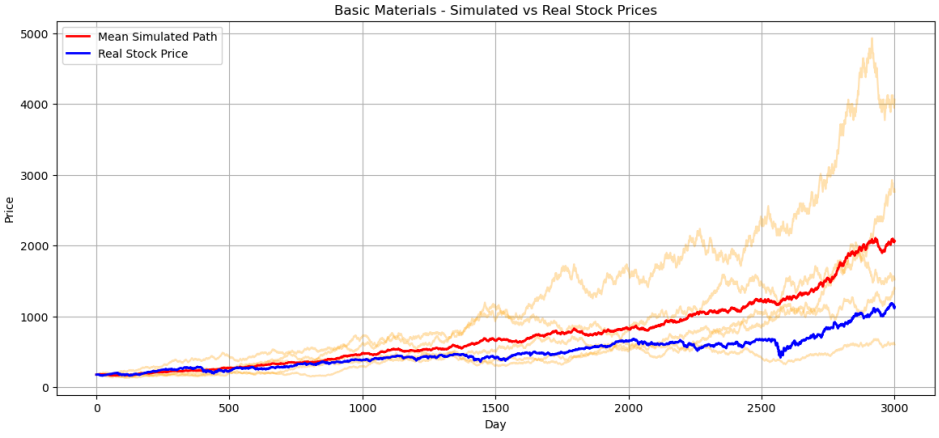

Industrials - Simulated vs Real Stock Prices
Mean Simulated Path
Real Stock Price
Price
Day
500
600
700
800
900
0
50
100
150
200
250
300
350
400

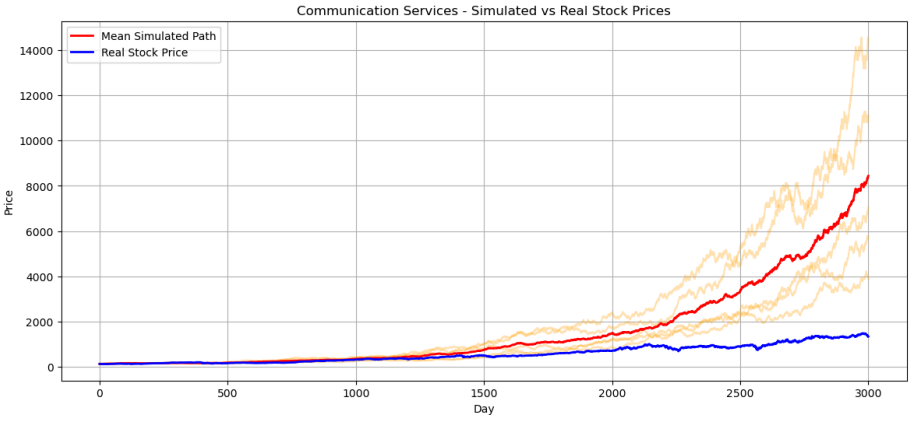

Real Estate - Simulated vs Real Stock Prices
Mean Simulated Path
Real Stock Price
Price
Day
250
300
350
400
450
500
550
0
50
100
150
200
250
300
350
400

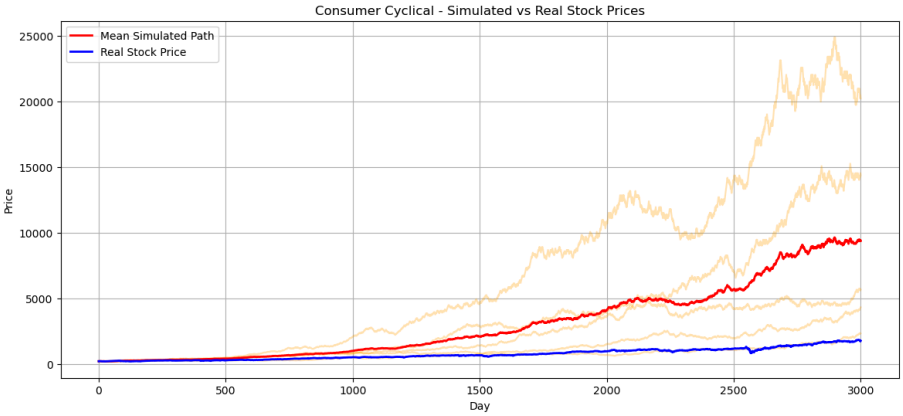


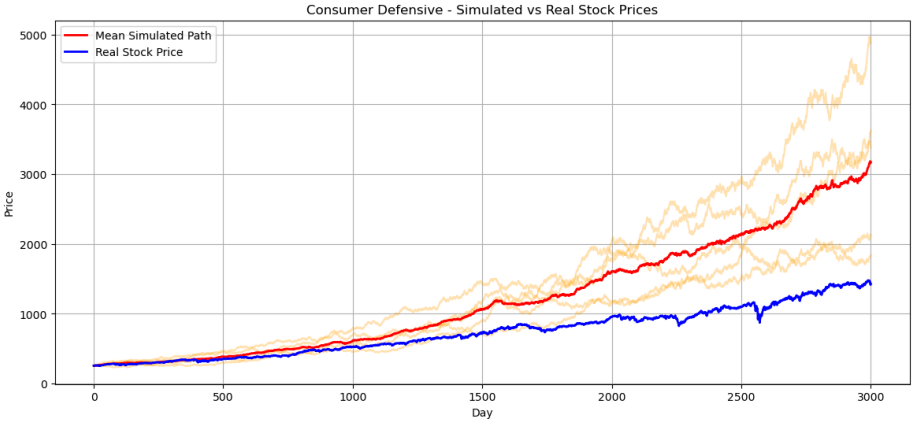


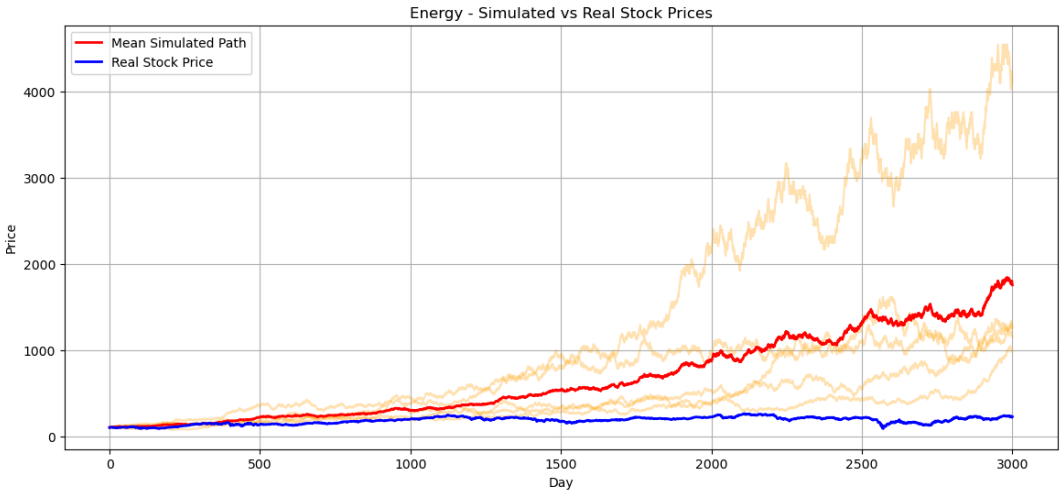
Long Term Model

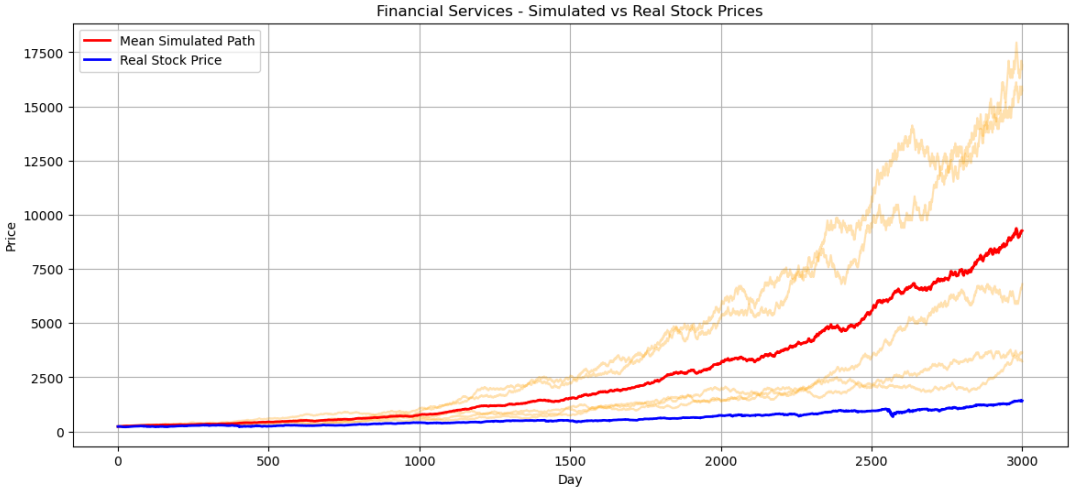


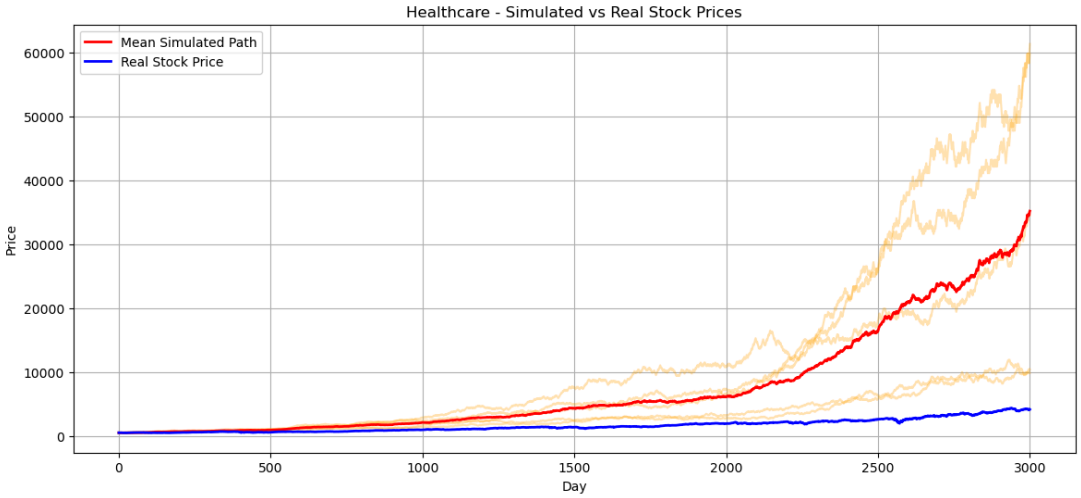

Communication Services - Simulated vs Real Stock Prices
Mean Simulated Path
Real Stock Price
Price
Day
0
2000
4000
6000
8000
10000
12000
14000
0
500
1000
1500
2000
2500
3000

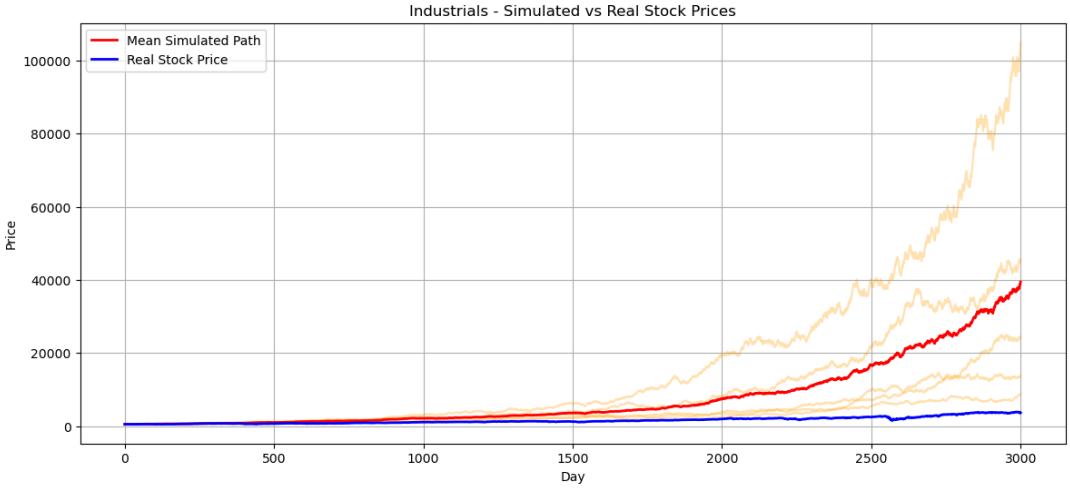

Consumer Cyclical - Simulated vs Real Stock Prices
Mean Simulated Path
Real Stock Price
Price
Day
0
5000
10000
15000
20000
25000
0
500
1000
1500
2000
2500
3000

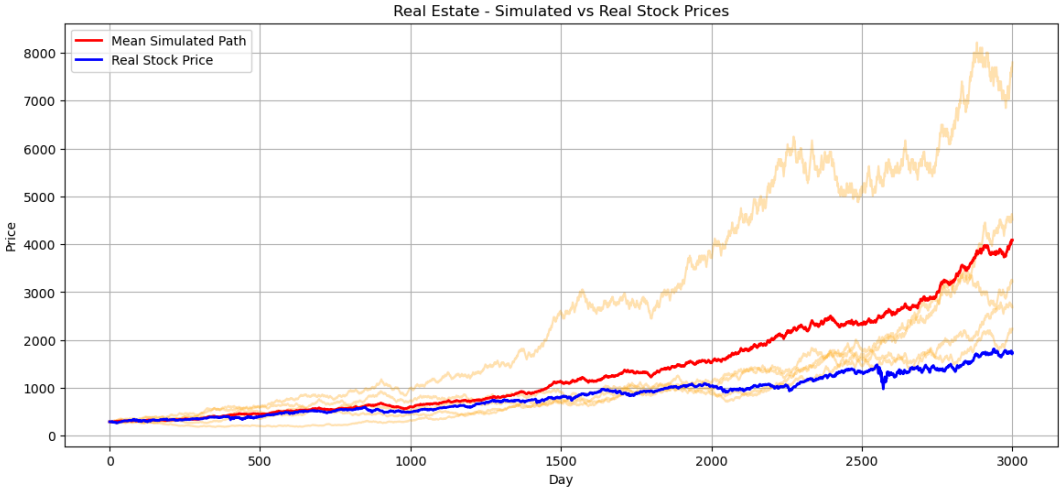

Consumer Defensive - Simulated vs Real Stock Prices
Mean Simulated Path
Real Stock Price
Price
Day
0
1000
2000
3000
4000
5000
0
500
1000
1500
2000
2500
3000

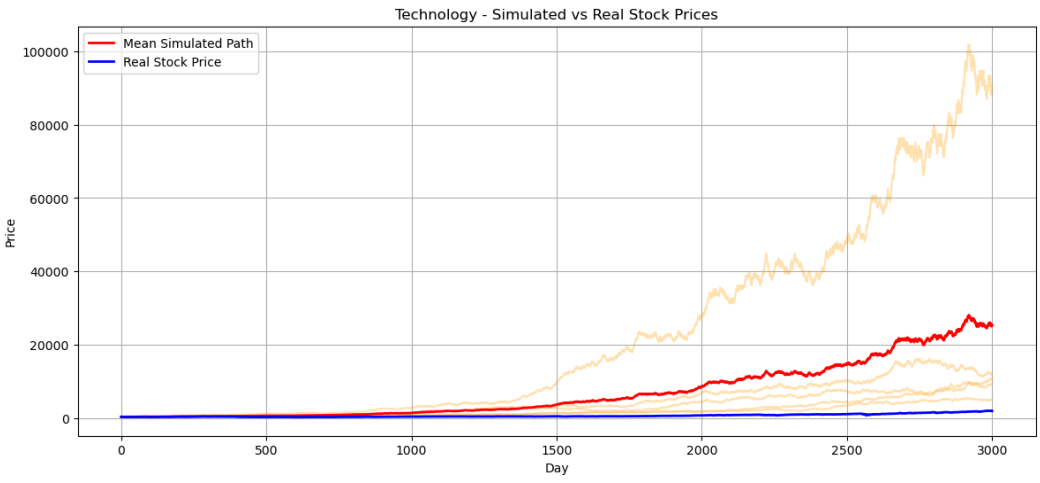

Energy - Simulated vs Real Stock Prices
Mean Simulated Path
Real Stock Price
Price
Day
0
1000
2000
3000
4000
0
500
1000
1500
2000
2500
3000

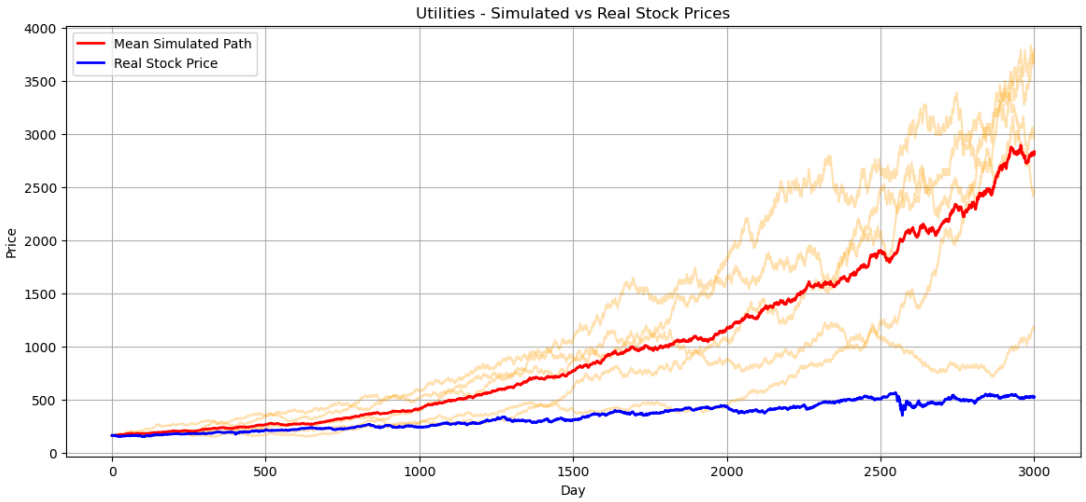

Financial Services - Simulated vs Real Stock Prices
Mean Simulated Path
Real Stock Price
Price
Day
0
2500
5000
7500
10000
12500
15000
17500
0
500
1000
1500
2000
2500
3000


Healthcare - Simulated vs Real Stock Prices
Mean Simulated Path
Real Stock Price
Price
Day
0
10000
20000
30000
40000
50000
60000
0
500
1000
1500
2000
2500
3000


Industrials - Simulated vs Real Stock Prices
Mean Simulated Path
Real Stock Price
Price
Day
0
20000
40000
60000
80000
100000
0
500
1000
1500
2000
2500
3000


Real Estate - Simulated vs Real Stock Prices
Mean Simulated Path
Real Stock Price
Price
Day
0
1000
2000
3000
4000
5000
6000
7000
8000
0
500
1000
1500
2000
2500
3000


Technology - Simulated vs Real Stock Prices
Mean Simulated Path
Real Stock Price
Price
Day
0
20000
40000
60000
80000
100000
0
500
1000
1500
2000
2500
3000



Model 2 Approach and Results

NOTE: All Charts/Code are viewable in ECDF_Random_Sample.ipynb

The objective of this model is to predict the value of a sector over time by generating new daily adjusted close values through random draws from the empirical distribution of daily percent changes. To extend these simulations over two future horizons, we applied the following procedure:

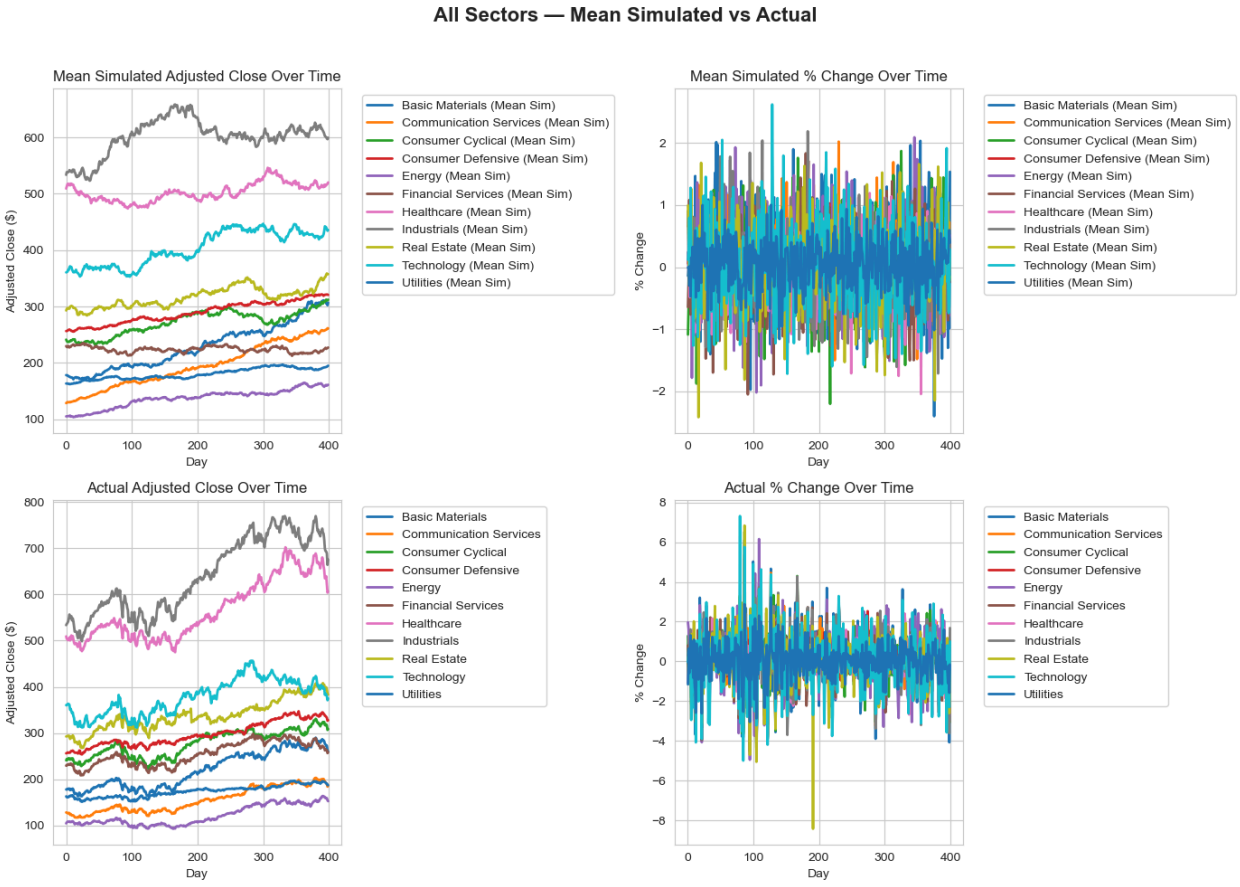
1. Begin on the second day of each sector, using the first day's adjusted close value as the starting point.
2. For each subsequent day, randomly draw a percent change from the existing empirical distribution of daily percent changes for that sector.
3. Compute the new adjusted close value by multiplying the previous day's value by *(1 + percent change / 100)*.
4. Repeat this process iteratively for each day until the desired forecast horizon is reached.

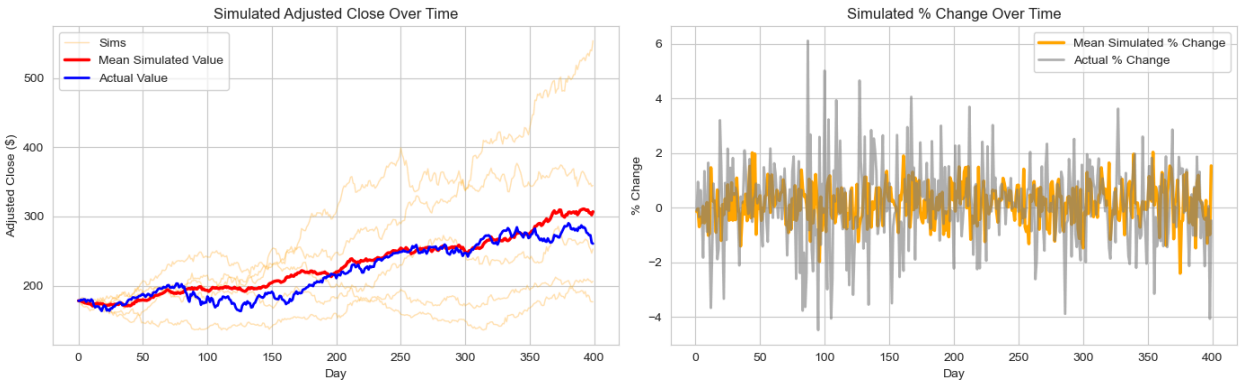
This approach ensured randomness in the sampling process, and running the simulation multiple times allowed for the aggregation of average model performance across each time horizon. It was selected because the dataset contained a large number of observations for each sector, and the empirical distributions (see *Models Overview*) were well-estimated. As a result, random sampling from these distributions provides a reliable and stable representation of the underlying population.

A key challenge in fitting this model was deciding how many simulated paths to generate and how to aggregate them. Averaging across too many simulations can smooth out the natural variability of the stock market, making the simulated trends appear unrealistically stable and

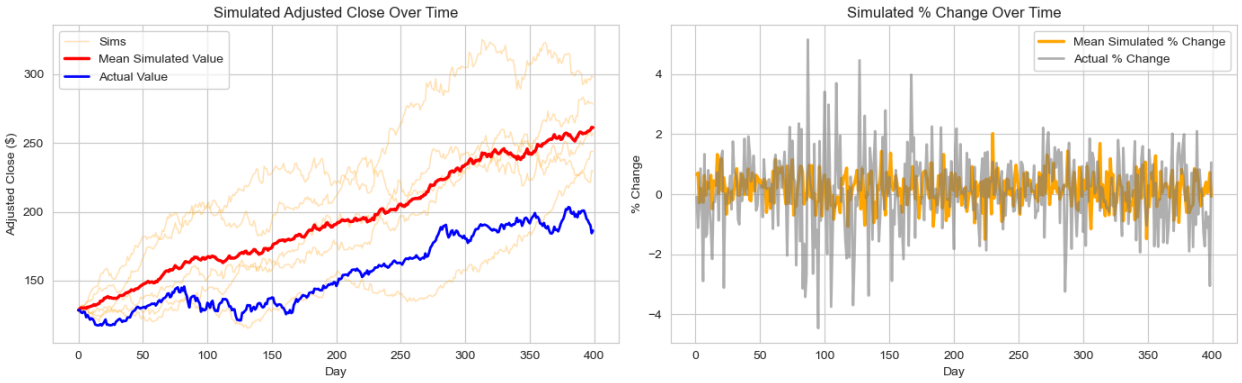
potentially underrepresenting real-world volatility. To preserve realistic variation while still obtaining stable estimates, a smaller number of simulations was chosen (5).

Short Term Model

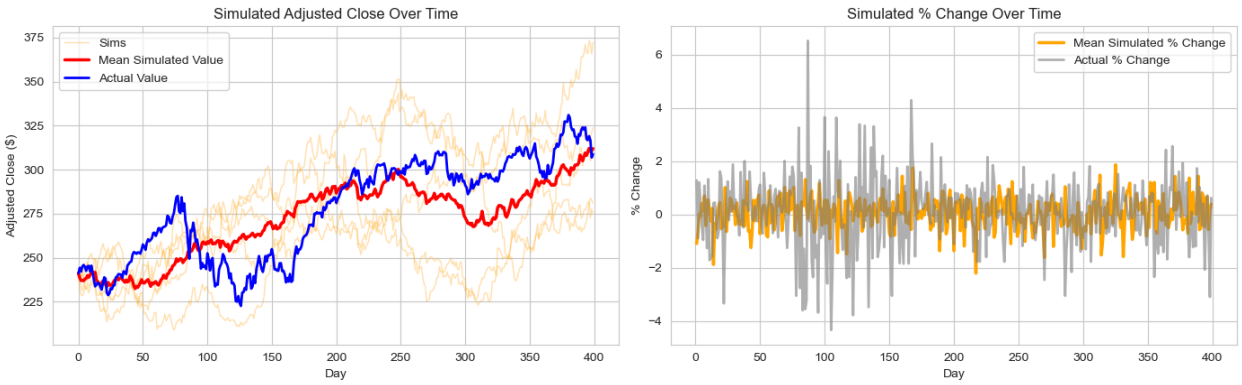




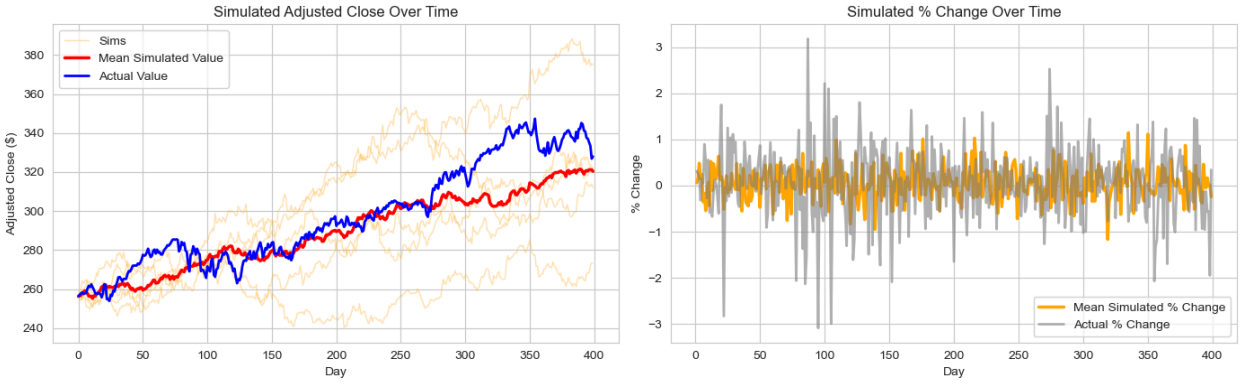
Communication Services Sector Simulation (5 Sims + Mean)



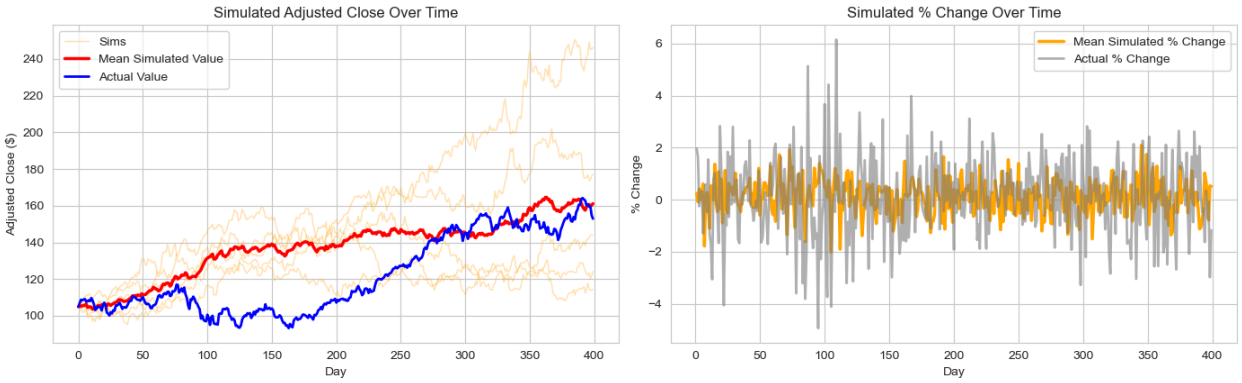
Consumer Cyclical Sector Simulation (5 Sims + Mean)



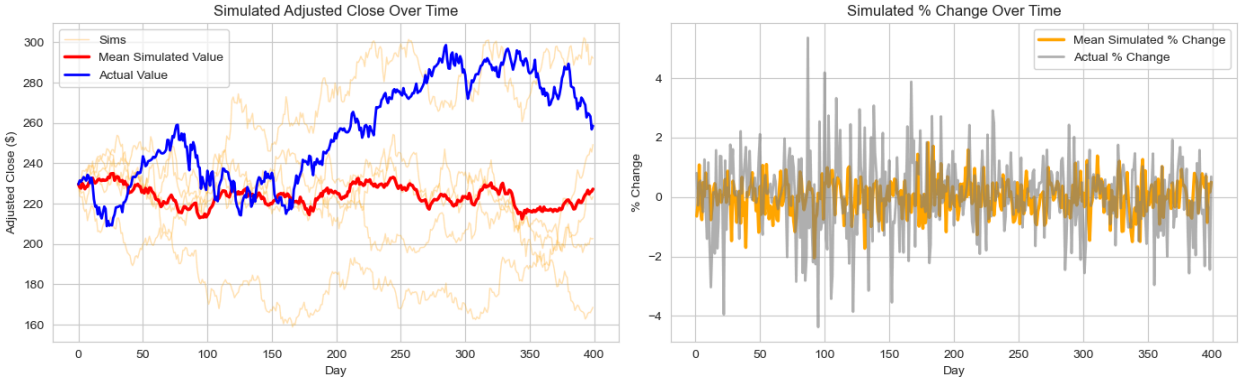
Consumer Defensive Sector Simulation (5 Sims + Mean)



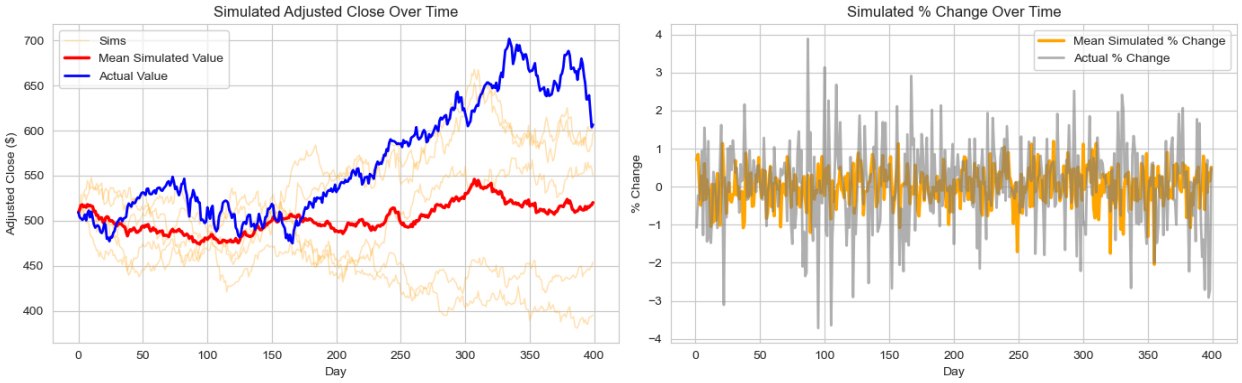
Energy Sector Simulation (5 Sims + Mean)



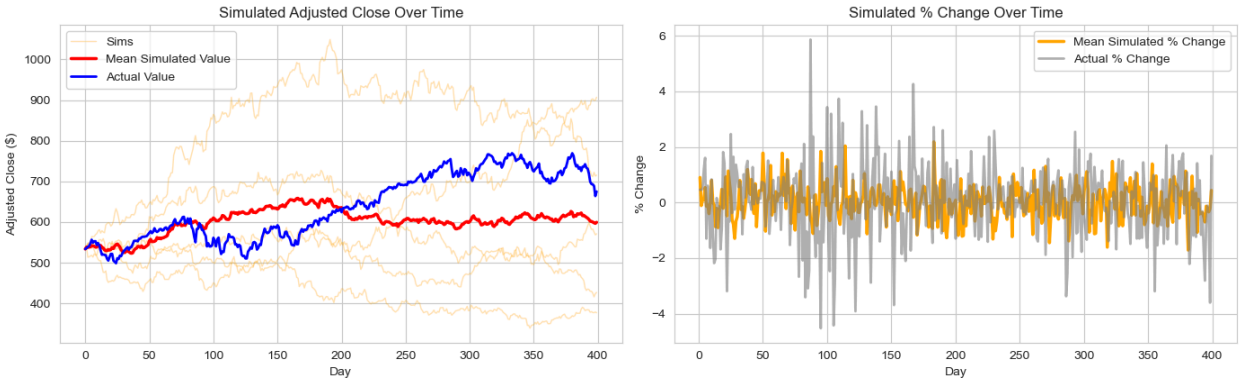
Financial Services Sector Simulation (5 Sims + Mean)



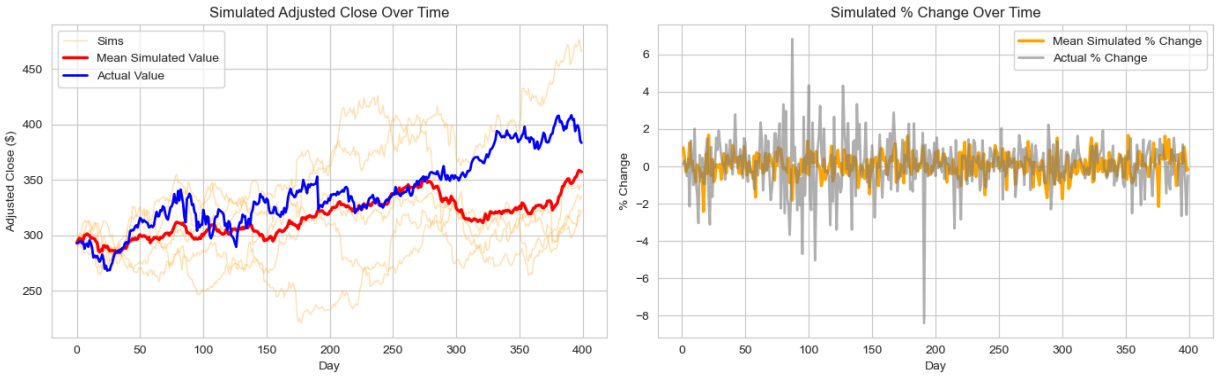
Healthcare Sector Simulation (5 Sims + Mean)



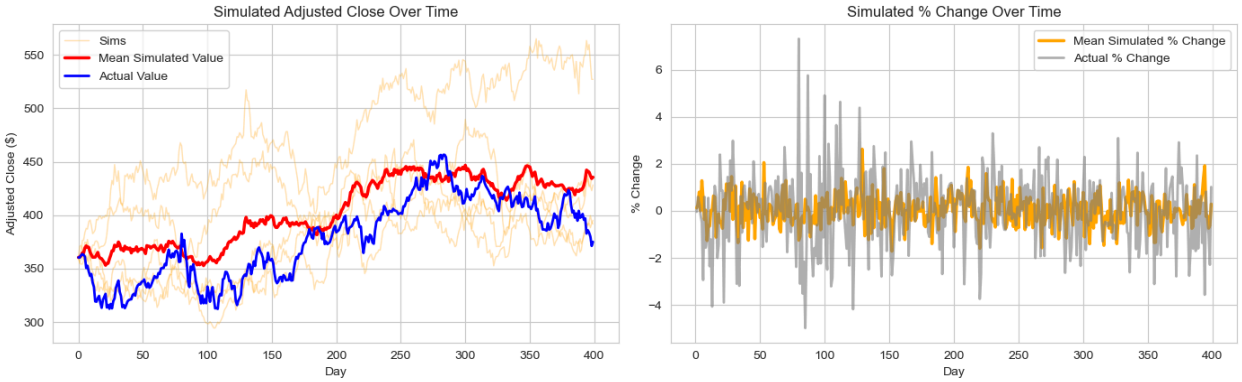
Industrials Sector Simulation (5 Sims + Mean)

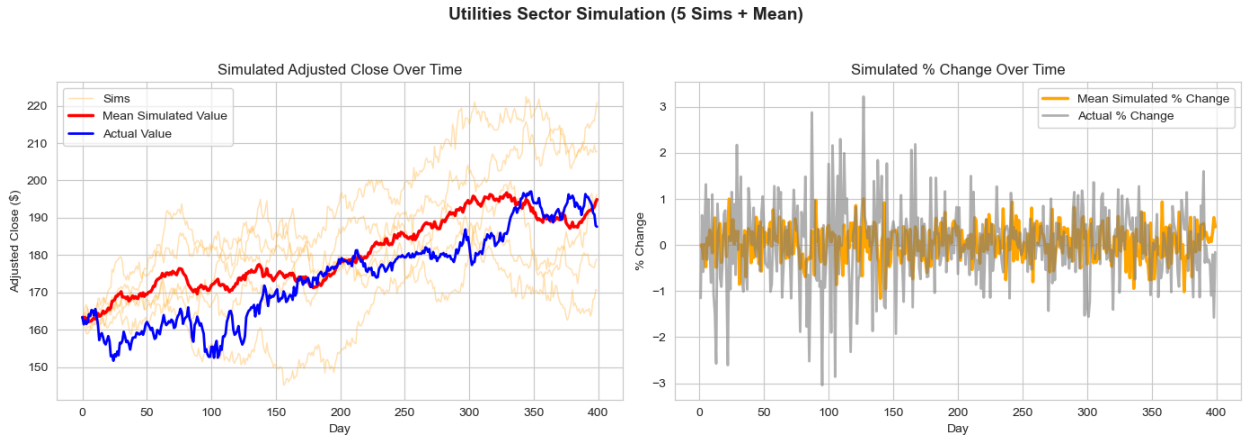


Real Estate Sector Simulation (5 Sims + Mean)

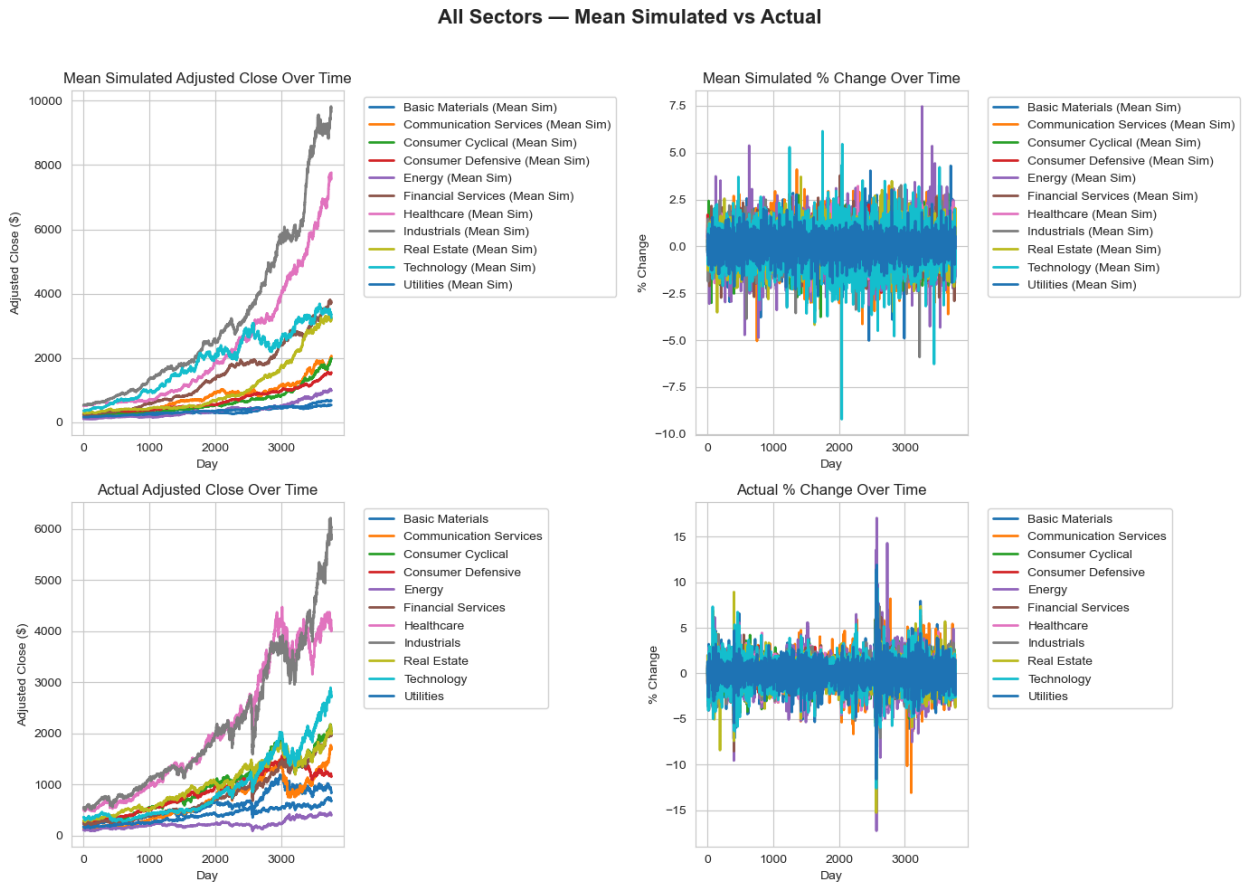


Technology Sector Simulation (5 Sims + Mean)

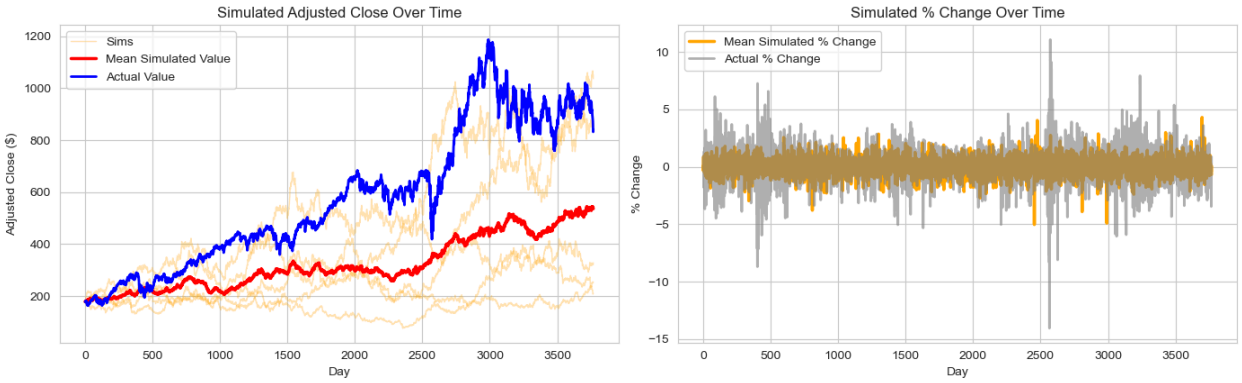




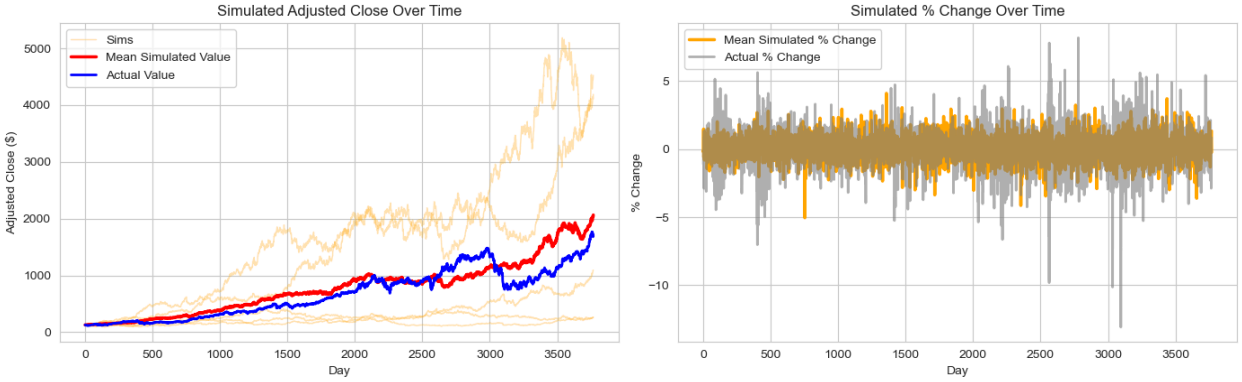
Long Term Model



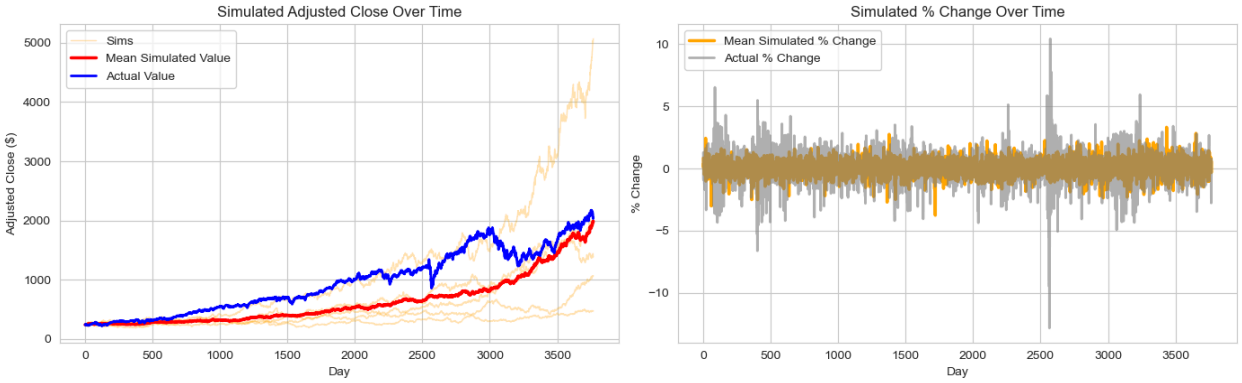
Basic Materials Sector Simulation (5 Sims + Mean)



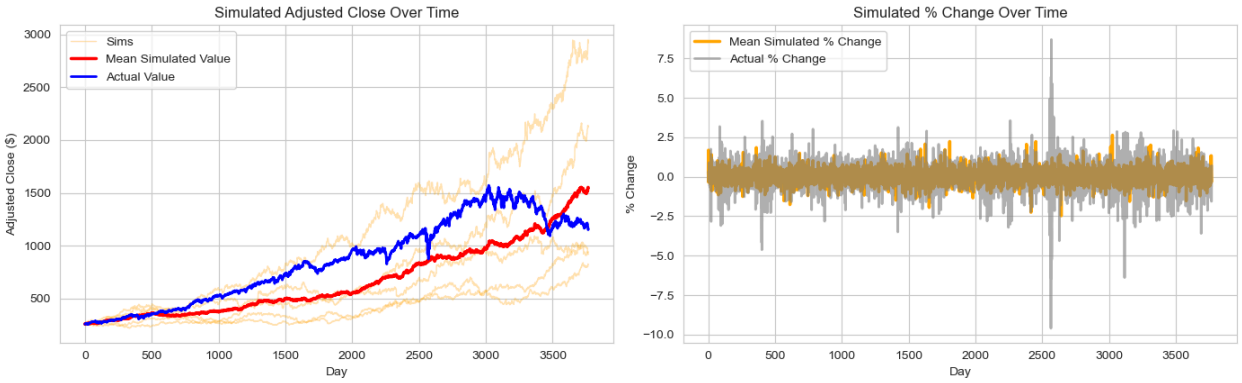
Communication Services Sector Simulation (5 Sims + Mean)



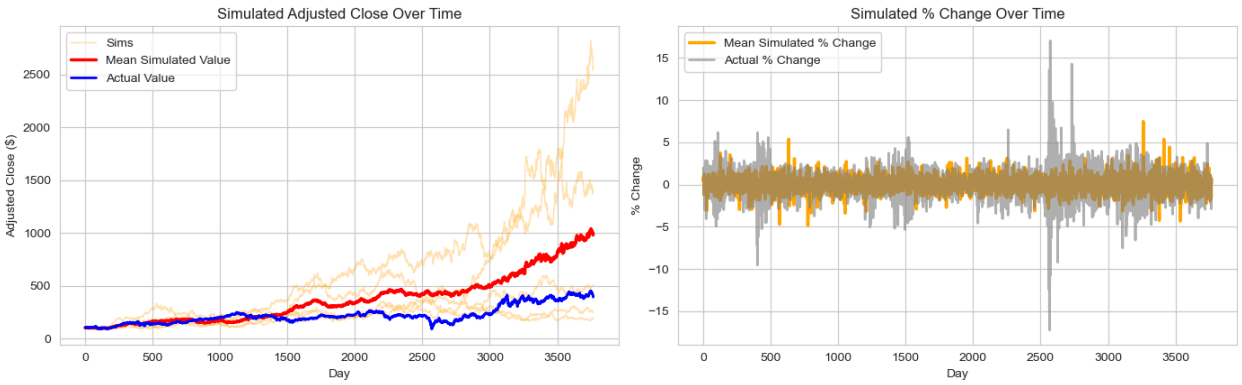
Consumer Cyclical Sector Simulation (5 Sims + Mean)



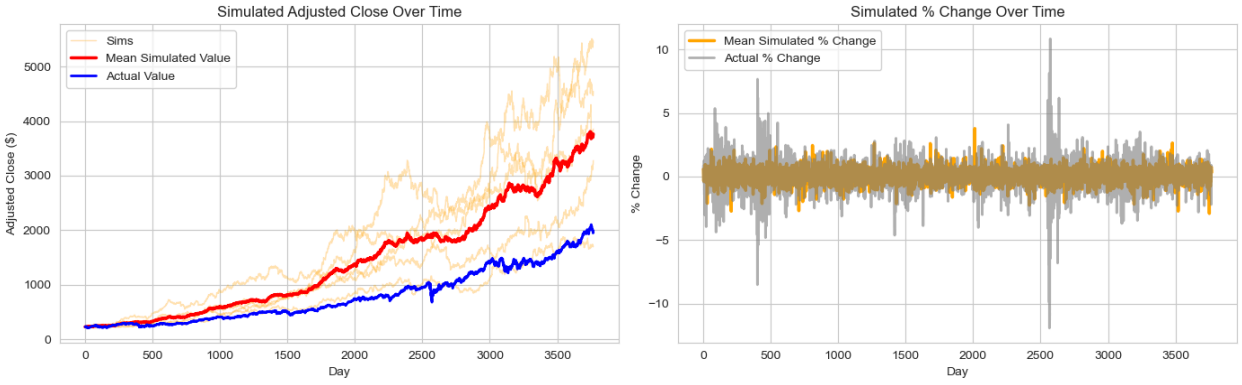
Consumer Defensive Sector Simulation (5 Sims + Mean)



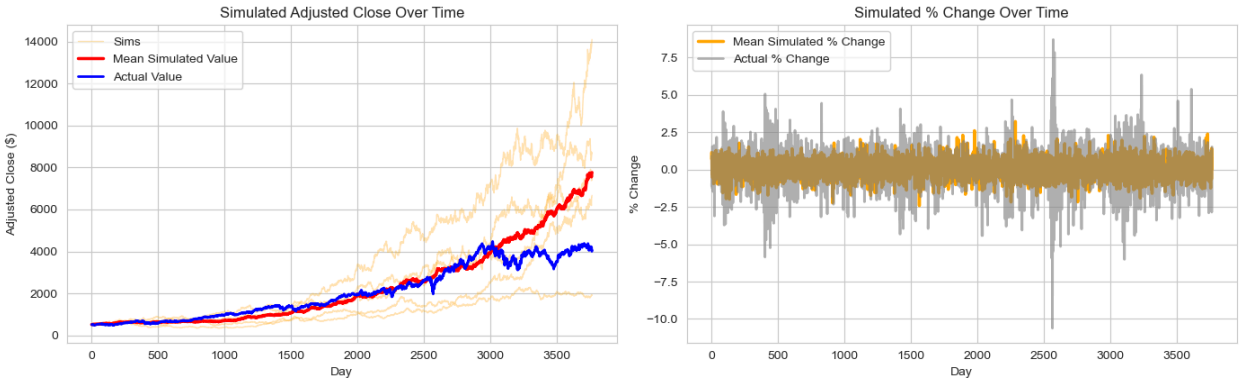
Energy Sector Simulation (5 Sims + Mean)



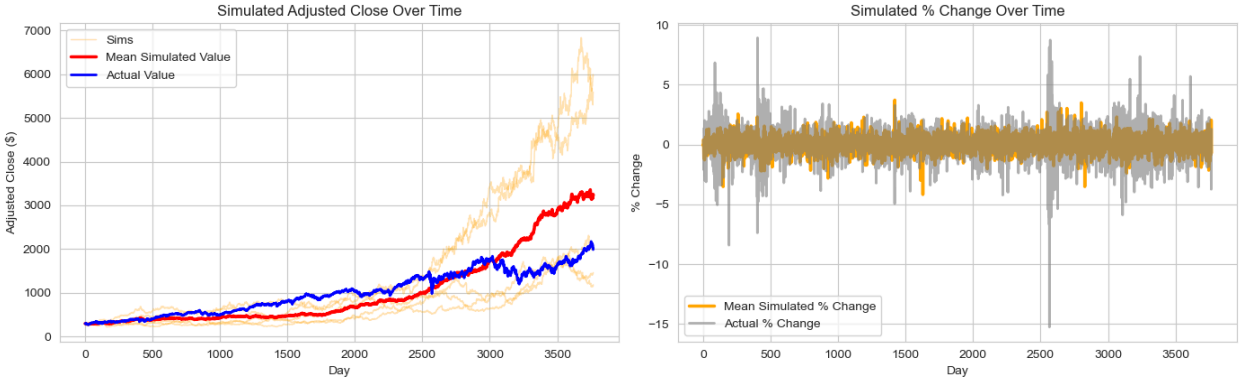
Financial Services Sector Simulation (5 Sims + Mean)



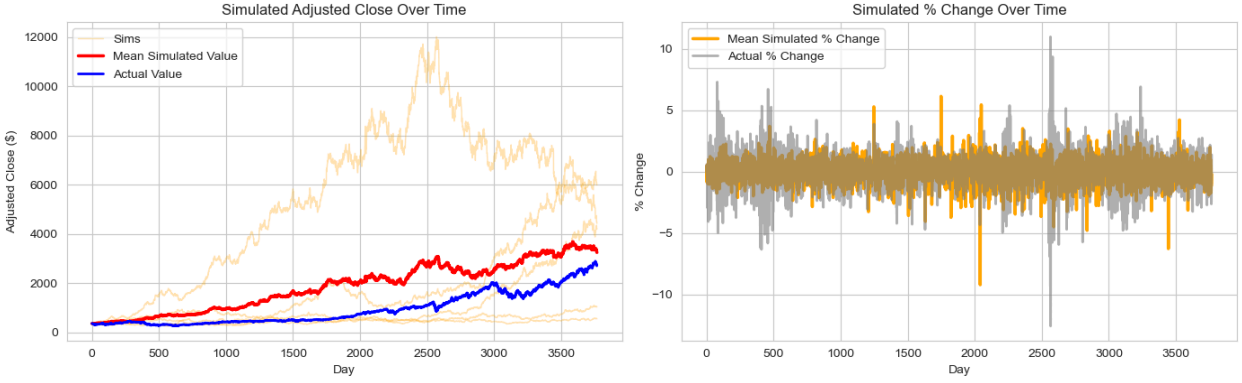
Healthcare Sector Simulation (5 Sims + Mean)

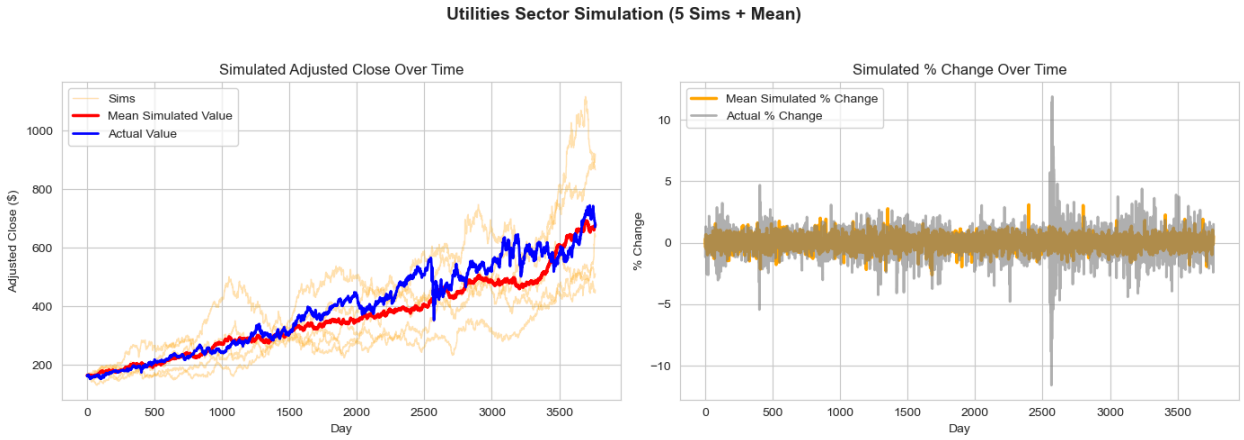


Real Estate Sector Simulation (5 Sims + Mean)



Technology Sector Simulation (5 Sims + Mean)





Cross Model Evaluation (Question 5)

NOTE: All Charts/Code are viewable in comparison_work.ipynb

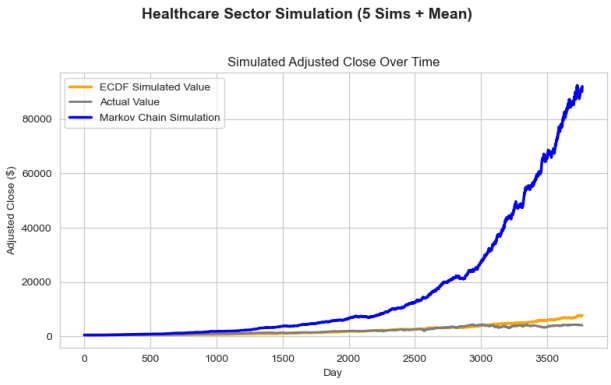
To evaluate our models developed in part 4, we compared the performance of the Markov Chain Simulation and the ECDF based simulation against actual stock closing values. This was conducted over the entire time period(around 3769) and then also for a shorter window of the first 400 days.

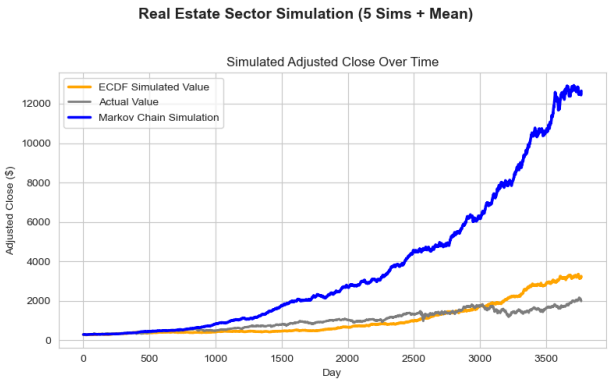
We focused on 3 sectors of interest for their perceived performance ‘Basic Materials’, ‘Healthcare’, and ‘Real Estate’ to best represent a range of volatility and growth patterns. We were then able to assess how each model performs across different sector dynamics.

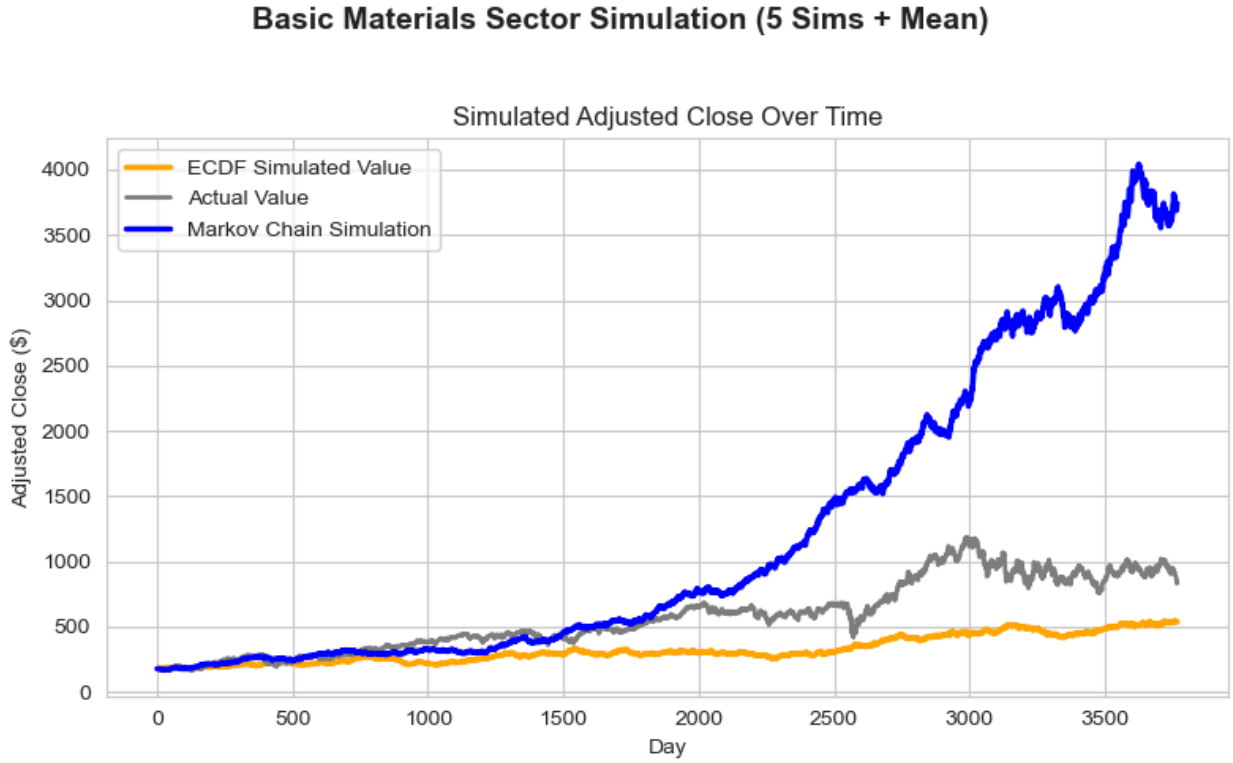
Performance Over the Full Time Period

The models showed markedly different behaviors at the full 3769 days. The Markov model diverged into excessively high estimates as the simulation progressed. In contrast, the ECDF model remained closer to actual values across most sectors, but still failed to fully capture periods of volatility.

On a more volatile sector like Basic Materials we see an underestimation by the ECDF simulation and an overestimation by the Markov model. In more constant sectors like Healthcare and Real Estate we see better performance from the ECDF at moderate accuracy and again extreme overestimation from the Markov model.







Graphs Representing the Two models Vs Actual for >3000 days

The reasons for this behavior can be explained by the nature of the models. As the number of days gets large the markov model approaches its steady state distribution. In this case the probability of increasing is greater than the probability of decreasing which combined with multiplying by a fixed volatility results in the exponential growth seen as the model explodes upwards overestimating the actual. This is seen every time showing the pattern of unreliability and demonstrates its unreliability in extended simulations.

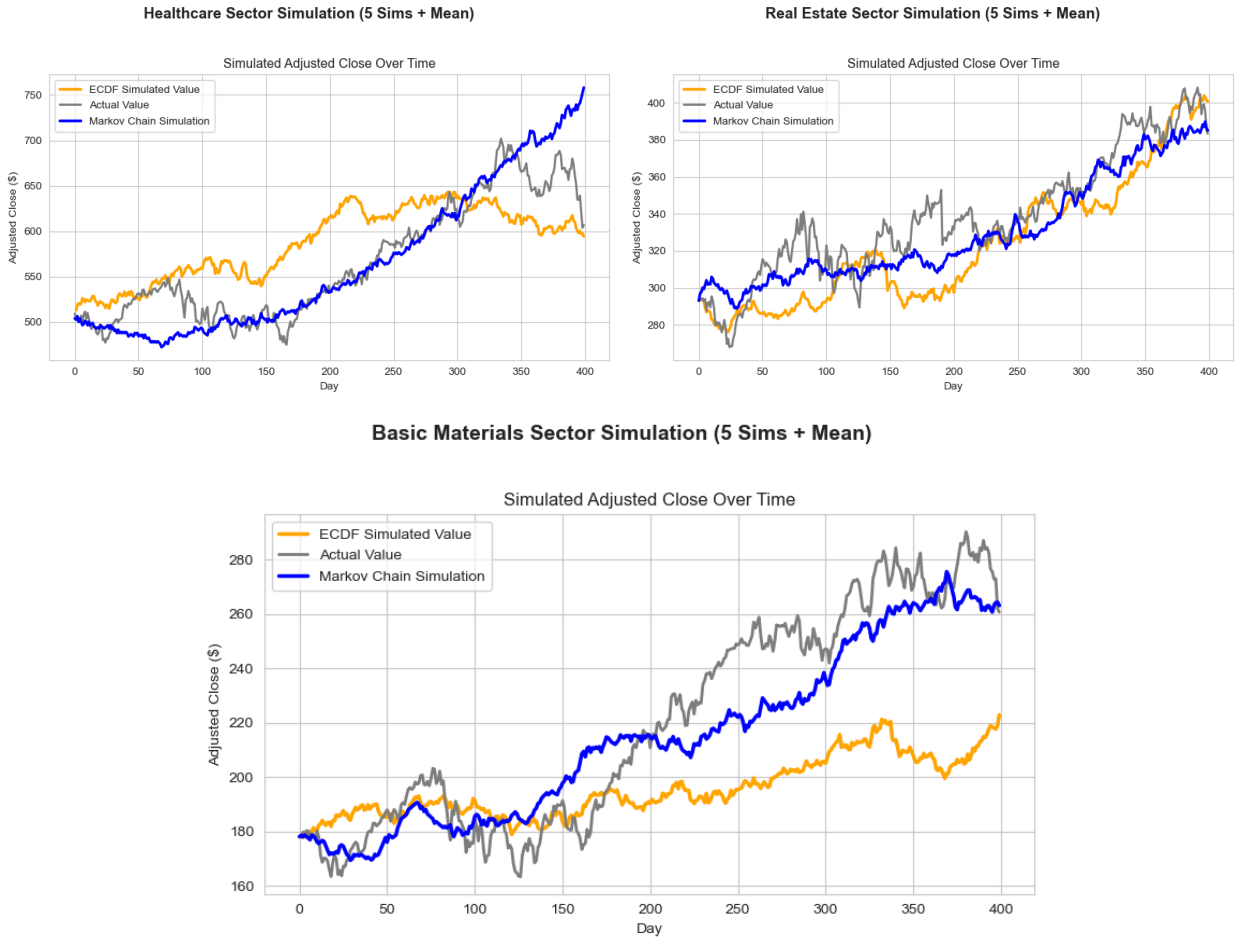
The ECDF model's performance stems from its reliance on random sampling. Because it relies on random sampling of the ECDF it is not capable of truly predicting major jumps in the sector's closing values. It is difficult for it to replicate extreme or infrequent jumps in value. As a result it performs somewhat reliably on more consistent sectors but struggles to capture the magnitude of the upward trend in the Basic Materials sector.

Overall neither model is an incredibly reliable long term predictor due to the limitations of their design. The Markov Chain overestimates growth and the ECDF ignores the volatility of some sectors.

Performance Over the First 400 Days

When evaluated over the much smaller window of 400 days the model performance tended to improve. The Markov model never reaches a steady state avoiding its extreme long term growth seen in the longer simulation. The ECDF model also benefits because a smaller window reduces the opportunities to be wrong, allowing it to maintain closer alignment while randomly sampling.

However neither approach fully captures the rapid shifts within the sectors. We can see with the Basic Materials chart that both the Markov Chain and the ECDF simulation underestimate the rapid growth it can experience, underscoring its unreliability.



Graphs Representing Markov vs ECDF Models Over 400 Days

Because rapid changes in a stock's closing value makes it difficult to predict its value, both the Markov model and the ECDF fail to consistently estimate the sector's final value. The Markov model appears to be the better predictor over the first 400 days for these sectors but more investigation is required to make a true evaluation.

Overall the performance is better in the short term for both models, but still struggle to truly predict the stock price over time.

Properties of Training Data and Reliability

Seeing the data in this way we can see that neither simulation model fully preserved the volatility and long term growth of the training data. Both models tend to mispredict sudden changes and the Markov Chain model overestimates when it begins to approach its steady state.

Given these faults, neither simulation is fully reliable. The ECDF model provides more stable and better long-term predictions but fails to account for rapid changes in the sector stock values. The Markov Chain offers a more volatile and energetic prediction option but still fails to encapsulate the changes in stock prices and fails to account for the steady state exponentiation of stock prices in the long term. Consequently uncertainty remains in our models regardless of the approach, highlighting the difficulty in fully encapsulating the real world changes in stock prices.

Limitations and Future Work (Question 6)

Our work on this topic has a few limitations. First off, we only included stocks in this analysis that were in the original dataset we used from kaggle and also had complete data for all the stock trading days from 01/04/2010 to 12/20/2024. This leads to our analysis not being as representative of real life stock sectors, as many of the sectors in our analysis only had 5 or more stocks used to create their Markov chain matrices and ECDFs. So, a solid first step to going further on this topic would be increasing the total number of stocks included in the analysis, as well as the total amount of days covered.

Limitations/Future Work Markov Specific:

While the Markov model does capture short-term momentum and patterns in price direction to some degree, it fails to consider the magnitude of changes. The extent to which a stock goes up or down one day would likely have some influence on how much it goes up or down the following day. In this way, our Markov chain based model reduces complex market behavior to only two discrete states (up/down). Additionally, we chose for simplicity to estimate volatility using the sample standard deviation of daily percentage changes in the stock price. This estimate is an empirical measure of historical volatility and fails to take into consideration future uncertainty. Going forward, a model that adjusts for volatility over time might be better at simulating changes in stock prices.

Limitations/Future Work ECDF Specific:

While the ECDFs of the stock sectors are useful for predicting stock price changes because it is based on the distribution of the percent changes, it has the limitation of not caring about the ordering of the percent change of a stock price. Most models used to predict stock price changes, including our Markov chains, factor in how a stock has been performing in recent times. The ECDF predictions do not care at all about what happened the day before, and are essentially treating the percent change of stock prices as independent, since the predictions from the ECDFs are based on taking a random percent change from the sector's percent change distribution. A potential change would be changing the random values that are being sampled from for each prediction, based on

what the last prediction was, as such to better account for the times in which a sector stock price might increase or decrease in a continued gradual manner. How the sample of the next predicted percent change would change based on previous predictions is not something we are sure how we would do, but might be interesting to explore when using actual distributions to make predictions.

We generally compared the models in our analysis visually, by plotting model predictions with the actual values of stock sector prices. Instead of a visual comparison, we could opt for a more quantitative one in future analysis on this topic, picking some kind of measure, like Mean Squared Error, to compare our model predicted values with the actual values. This might allow greater specificity and understanding the magnitude for which our models work (or don't work).

While this project centered on predictions using Markov chains and ECDFs to generate predictions for stock sectors, we could explore other methods like more traditional time series models that are used for predicting the changes to stock price. Using the data we have, we can compare a number of additional modeling methods to get a better idea of what models/techniques are actually most effective in understanding the stock market.

Even with any limitations and other areas that could be explored later, we still gained a certain understanding of how stock sectors perform and vary. We are able to see that Markov chains are more useful than ECDFs when it comes to predicting stock performance over a smaller timeline, while when the timeline of prediction is expanded ECDFs tend to perform better than Markov chains. We can see that regardless of stock sectors, growth is predicted in all of them, with some sectors predicted slightly more often to increase on a random day than others.