Instituto Tecnológico y de Estudios Superiores de Occidente

Reconocimiento de validez oficial de estudios de nivel superior según acuerdo secretarial 15018, publicado en el Diario Oficial de la Federación del 29 de noviembre de 1976.

Department of Mathematics and Physics
Master of Data Science

Optimized pathway for non-experienced investors

THESIS to obtain the **DEGREE** of
MASTER OF DATA SCIENCE

A thesis presented by:
Rosa María Sánchez Leguízamo

Thesis Advisor:
Dr. Luis Raúl Rodríguez Reyes

Tlaquepaque, Jalisco, November 29, 2024

Instituto Tecnológico y de Estudios Superiores de Occidente

Reconocimiento de validez oficial de estudios de nivel superior según acuerdo secretarial 15018, publicado en el Diario Oficial de la Federación del 29 de noviembre de 1976.

Department of Mathematics and Physics

Master of Data Science Approval Form

Thesis Title: **Optimized pathway for non-experienced investors**

Author: **Rosa María Sánchez Leguízamo**

Thesis Approved to complete all degree requirements for the Master of Science Degree in Data Science.

Thesis Advisor, **Dr. Luis Raúl Rodríguez Reyes**

Thesis Reader, **Dr. Alma Nayeli Rodríguez Vázquez**

Thesis Reader, **Dr. María del Rosario Ruiz Hernández**

Tlaquepaque, Jalisco, November 29, 2024

Optimized pathway for non-experienced investors

Rosa María Sánchez Leguízamo

Abstract

Currently, we have multiple ways (brokers, applications, portfolios and tickers) to start trading, various limitations (minimum time to invest, minimum investment amount, open market schedules, currencies), multiple costs to open the account, administer it, commissions, taxes and multiple risks (index changes, portfolio weight changes, conversion currency loss) among others. In summary, choosing the optimal option for the specific requirements is difficult for non-experienced investors (anybody with zero knowledge of financial markets or new to trading who relies on external advice or emotions to make decisions).

We will apply an optimization process to trading to create a portfolio between 15 to 30 tickers (diversification) to then select the most efficient scenario to maximize the returns, minimize the risk and reduce the costs by trading at a strategical timing to help non-experienced investors managing their investments effectively.

There are some difficulties when talking about trading: emotions (fear or greed) which can lead to impulsive actions, unknown risk appetite which can lead to taking more risk than necessary, stop-loss (S/L) to limit the losses and take-profit (T/P) to lock in profits.

On the other hand, some other risks are associated with the interaction between the market and the optimisation process. It evolves continuously and considerably so the optimisation process should constantly adapt to changing conditions. Finding the right balance between performance and robustness with past and present data (historical data) can work perfectly in theory but poorly in real-time, leading to associated risks and costs.

The preliminary results of 85% successful rate for variable versus static scenarios and an extra $ 64,944 USD gained (37% of $ 100 KUSD investment) from 34 of 40 tests provide another opportunity for non-experienced investors to get into the trading world without paying unnecessary risks and/or high administrative costs or being forced to select only from the preselected portfolios available through brokers and rigid periods.

Ruta optimizada para inversionistas sin experiencia

Rosa María Sánchez Leguízamo

Resumen

Actualmente, exiten muchas formas (corredores de bolsa, aplicaciones, portafolios y acciones) de comenzar a invertir, varias limitaciones (tiempo mínimo de inversión, monto mínimo, horarios específicos y monedas), multiples costos para abrir una cuenta, administrarla, comisiones, impuestos y riesgos financieros (cambios de índices, cambios en distribución de compañías de los portafolios, pérdida por tipo de cambio) entre otras. En resumen, elegir la mejor opcón para las necesidades de cada inversionista sin experiencia (cualquier persona con cero conocimiento en inversiones o nuevo, que aun depende de personas externas y toma decisiones basado en las emociones) es difícil.

En el mundo de trading existen algunas dificultades: emociones (miedo o codicia) que nos hacen tomar acciones impulsivas o arriesgarnos más de lo necesario por desconocimiento tratando de limitar las pérdidas conocido como stop-loss (S/L) o tratando de asegurar las ganancias conocido como take-profit (T/P).

Para este trabajo, aplicaremos un proceso de optimización para crear portafolios de inversión entre 15 a 30 activos (diversificación) para seleccionar el portafolio óptimo que maximice el retorno de inversión, minimice el riesgo y reduzca el costo invirtiendo en un tiempo estratégico de acuerdo con las necesidades de un inversionista sin experiencia en el mercado.

Por otro lado, algunos otros riesgos están asociados con el comportamiento entre el mercado y el proceso de optimización. Este evoluciona continuamente y considerablemente por lo que el proceso de optimización deberá adaptarse de igual manera considerando siempre el balance entre rendimiento y robustez debido a que los datos históricos y actuales pueden funcionar muy bien en teoría y pobremente en tiempo real lo que podría generar riesgos y costos adicionales.

La tasa de éxito del 85% (34 de 40 pruebas) y la ganancia adicional de $ 64,944 USD obtenidos de escenarios variables por encima de

escenarios estáticos (37% de $ 100 KUSD de inversión) como resultados preliminares brindan otra oportunidad para que los inversionistase inexpertos incursionen en el mundo del trading sin pagar riesgos innecesarios y / o altos costos administrativos o verse obligados a seleccionar solo de las carteras preseleccionadas disponibles a través de corredores y períodos rígidos.

Contents

List of Figures

List of Tables

Dedicated to ....

All who have walked with me on this challenging but fulfilling journey, particularly those who in one way or another contributed and extended their valuable assistance in the preparation and completion of this study.

Firstly, I would like to express my sincere thanks to the academic staff of Instituto Tecnológico y de Estudios Superiores de Occidente (ITESO) University who provided their invaluable expertise and guidance. Special appreciation to my supervisor **Dr. Luis Raúl Rodríguez Reyes**, whose unwavering guidance, experience, suggestions and support have been instrumental in the completion of this thesis. I extend my deepest gratitude to **Sergio Tinoco** who has shared his experience and perspective to enrich the overall development and results of this paper.

Special thanks go to **my family**, whose constant encouragement fueled my perseverance during the completion of this dissertation. Finally, my heartfelt gratitude goes out to my friends **Gilberto Ponce**, **Mariana Urbieta** and **Cinthya Guerra**, who have been my solid pillars of strength and consistently believed in my potential for growth.

Finally, it's a pleasure sharing these pages with the readers. I hope their journey through these pages is enjoyable and sparks a curiosity to keep swimming into the vast, uncharted ocean of discovery. Wishing them a wonderful life and until our paths cross again... farewell.

Dedicado a ....

Todos los que han caminado conmigo en este desafiante pero gratificante viaje, particularmente aquellos que de una forma u otra contribuyeron y brindaron su valiosa asistencia en la preparación y finalización de este estudio.

En primer lugar, me gustaría expresar mi más sincero agradecimiento al personal académico del Instituto Tecnológico y de Estudios Superiores de Occidente (ITESO) que brindó su invaluable experiencia y orientación. Un agradecimiento especial a mi director **Dr. Luis Raúl Rodríguez Reyes**, cuya invaluable guía, experiencia, sugerencias y apoyo han sido fundamentales para la realización de esta tesis. Extiendo mi más profundo agradecimiento a **Sergio Tinoco** quien ha compartido su experiencia y perspectiva para enriquecer el desarrollo general y los resultados de este contenido.

Un agradecimiento especial a mi **familia**, cuyo constante aliento impulsó mi perseverancia durante la realización de este trabajo de obtención de grado. Finalmente expreso mi más sincero agradecimiento a mis amigos **Gilberto Ponce**, **Mariana Urbieta** y **Cinthya Guerra** quienes han sido mis pilares de fortaleza y siempre han creído que puedo convertirme en una mejor versión de mí.

Finalmente, es un placer compartir estas páginas con los lectores. Espero que su viaje a través de estas páginas sea agradable y despierte la curiosidad de seguir nadando en el vasto e inexplorado océano del descubrimiento. Deseándoles una vida maravillosa y hasta que nuestros caminos se vuelvan a cruzar... adiós.

1 Introduction

This document is described as simply as possible considering our main readers are non-experienced investors. However, as in almost every topic, we can find plenty of literature on the different areas we will be describing in this document for which we suggest as a good source of definitions the Investopedia website [1]: trading, portfolio, tickers, diversification, maximize the returns, minimize the risk, stop-loss (S/L) and take-profit (T/P).

[1] Investopedia. Investopedia, Year the website was last updated. URL https://www.investopedia.com/. Accessed on Date

We will demonstrate the net gains (including the associated fees paid) are higher or the net losses are lower when optimizing the portfolio and frequently trading over a period (we will call this variable scenario) than when optimizing the portfolio and only trading once during the same determined time (we will call this static scenario) even when using the most reasonable stock price (open vs close) for each scenario.

We will be running the optimization[2] and trading process multiple times using different periods as well as using multiple historical data from the same data source and keeping the same entry data (investment, number of tickers in the portfolio) capturing the results and then comparing the results and adjusting as necessary.

[2] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, 2004

We will describe in detail all the platforms and data used considering this is a capstone project for a Data Science degree. As a future work, if this goes into the market, it will be mandatory to create a friendly application for the non-experienced investor to capture the initial values (configuration/parameters), show the results, provide historical data and easily interact with the final user.

There is absolutely no intention for the non-experienced investor to have to configure everything included in this document nor go into all the details described; as they are expected to only interact with a mobile application. However, we expressed everything simple so anybody can enjoy the reading and possibly awaken their curiosity to develop this further.

2 Platforms and data

Contents

To ensure the accuracy of our success, it is crucial to consistently use the same data, parameters, and optimization process when running the process for both scenarios (static and variable).

Let's talk about the different components we have used to run the tests and how each is relevant, can be changed or adjusted if necessary and if there is any particularity we need to consider when running them.

2.1 Platforms

The first component is the platforms, 2 of them allow us to collect the data and eventually run a real-time test if necessary while the 3rd one allows us to classify, process and transform our data as required to then apply the optimization process and finally run the tests to gather results, compare them and calculate the successful rate; so, let's review each of them in detail.

2.1.1 TradingView

TradingView[1] is an analytical platform and social network for traders

[1] Tradingview: Free stock charts, stock quotes and trade ideas. https://www.tradingview.com/, a. Accessed on Date

and investors to explore, analyze and share market insights. You can find its definition on its website[2]. This platform doesn't offer on its own a mechanism to download the data nor an API; however, it allows certified brokers to get connected for real-time (automatic/programmed) transactions. All the information regarding the API can be consulted in the FAQ widget documents reference page[3].

The platform is recommended for day and swing traders and for this research, the free license can provide the necessary information and the only requirement to have access to it is to register. Nevertheless, as we can see in the table 2.1 below the platform offers much more functionalities for different prices[4]. Stock Analysis provides the most completed review on TradingView for experienced traders[5].

Although TradingView can offer years of historical data and the actual tests have proven to bring up to 2 years of data, we will only collect between 60 to 120 days for results and success rate comparison; however, you can review all the details about the historical data in the how-to page[6].

[2] Tradingview: Definition. https://en.wikipedia.org/wiki/TradingView, c. Accessed on Date

[3] Tradingview data download and api information. https://www.tradingview.com/widget-docs/faq/data/, b. Accessed on Date

[4] Tradingview price information. https://www.tradingview.com/pricing/, d. Accessed on Date

[5] StockAnalysis. Tradingview review. https://stockanalysis.com/article/tradingview-review/. Accessed on Date

[6] TradingView. How to see the deepest historical data. https://shorturl.at/yahJg. Accessed on Date

Table 2.1: TradingView list of USD prices with features (July 2024)

Feature	Free	Essential	Plus	Premium
Cost(monthly)	$ 0.00	$ 12.95	$ 24.95	$ 49.95
Charts per layout	1	2	4	8
Indicators per chart	2	5	10	25
Price alerts	5	20	100	400
Ads	YES	NO	NO	NO
Bar reply	YES	NO	NO	NO
Watchlists	1	Unlimited	Unlimited	Unlimited

2.1.2 *Alpaca*

As per their site[7] Alpaca[8] is a global technology company that provides a commission-free stock trading API that allows real-time price, trades, manages the portfolio and streams for advanced financial investors as a member of FINRA/SIPC.

Alpaca will also require the user to create an account and fill in all necessary legal and tax documents to be able to start trading. Once the account is created the platform will provide $ 100,000 USD in paper which can be used for testing purposes and we can also wire money to start trading for real.

Alpaca as a certified broker for TradingView and can be added to the Trading Panel by logging in with your Alpaca credentials so the

[7] Alpaca. What is alpaca. https://alpaca.markets/support/what-is-alpaca, c. Accessed on Date

[8] Alpaca. Alpaca markets. https://app.alpaca.markets/, b. Accessed on Date

investors can see their bought stocks in real-time and any of the other TradingView functionalities included in their license.

More importantly, the Alpaca API [9] allows us to bring the historical data from TradingView by configuring and calling the API in any programming language that supports API connections.

[9] Alpaca. Alpaca api platform. https://docs.alpaca.markets/docs/alpaca-api-platform, a. Accessed on Date

2.1.3 *Python*

As per their site Python[10] is an easy-to-read syntax and dynamic typing high-level programming language. It is a simple and trendy programming language, widely used and adopted due to its vibrant community producing useful libraries that facilitate data analysis, artificial intelligence, scientific computing and automation development work.

[10] Python Software Foundation. Python: The python programming language. https://www.python.org/. Accessed on Date

It has become one of the most popular programming languages due to multiple qualities: easy to learn, scalable, wide variety of libraries, excellent to manage data, well supported and easy to integrate with other tools. See more information in the article: "What makes Python a brilliant choice for data analysis"[11].

[11] P. K. Jha. What makes python a brilliant choice for data analysis. https://shorturl.at/09zgb. Accessed on Date

Python will allow us to collect the data, curate it, apply the optimization process, gather the results and finally measure the success rate very easily.

There is a wide variety of websites that talk about the benefits of using Python for Data Science as well as the most common programming languages for this sort of problem for which we reference another very complete article: "Top 10 Best Programming Languages To Learn in 2024"[12].

[12] SuperiorCodeLabs. Top 10 best programming languages to learn in 2024. https://shorturl.at/bmHvr. Accessed on Date

2.2 *Data*

The second component is the data, there are multiple factors we need to consider when talking about trading. Each of those factors changes how we need to analyze the data and the interpretation of the results. We will be describing in detail each of the factors we have selected and how we will be using and interpreting them. In addition to that, we have also included information on how these can be adjusted and the suggestions to be considered if necessary.

2.2.1 *Market indexes*

As per Michele Cagan [13][p. 84], a stock index is a group of independent stocks based on distinct of their features. Some of them are focused on specific sectors while others keep a balanced selection of different sectors. We can even choose and compare some of these indexes to understand how well or poor a particular stock is doing.Three of the most common stock indexes in the United States are the Dow Jones Industrial Average (the Dow) [14], the S&P 500, and the NASDAQ Composite. We can find a great variety of indexes available and we can decide which of them could be more aligned to what we are looking for.

[13] M. Cagan. *Stock Market 101: From Bull and Bear Markets to Dividends, Shares, and Margins—Your Essential Guide to the Stock Market*. Adams Media, kindle edition edition, 2016

[14] Wikipedia. Dow jones industrial average, 2024a. URL https://en.wikipedia.org/wiki/Dow_Jones_Industrial_Average. [Online; accessed 30-May-2024]

For this research, it is important for us to select stable or growing indexes that allows us to run multiple tests over time and be able to get results we can still compare easily. We also want to select indexes which have between 30 to 50 tickers so we can run our tests with plenty tickers to select from that are somwhow similar. So, we will use the 30 tickers from the Dow Jones index (DJIA) [15] for which we can see how it has got stronger over time from 1896 to 2024 in figure 2.1, then see the behavior over the last 5 years in figure 2.2, then over the last year in figure 2.3 and finally its behavior year to November 25th in figure 2.4.

[15] TradingView. Dji, 2024. URL https://www.tradingview.com/symbols/TVC-DJI/. [Online; accessed 30-May-2024]

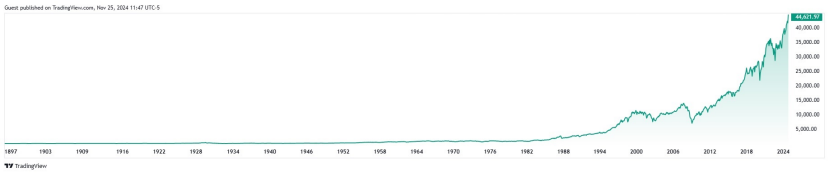

Figure 2.1: TradingView Dow Jones Industrial 1986 to 2024. *n. Select "All time" tab at the website for current data.*

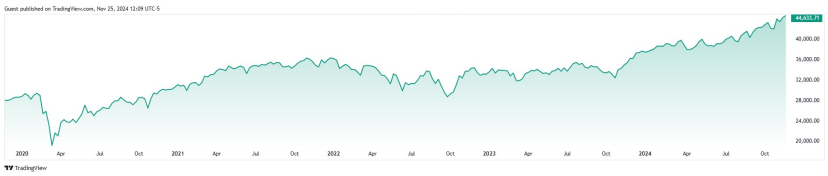

Figure 2.2: TradingView Dow Jones Industrial last 5 years. *n. Select "5 years" tab at the website for current data.*

The growing condition of this index will consequently show the prices of the tickers have also increased and as such the calculated weights in the optimal portfolio will be impacted.

Let's review in detail table 2.2 of companies included in the Dow Jones index, the ticker symbol which will be used when extracting the

data as well as how each of these companies is distributed (represented by weights) within the Dow Jones Industrial index. We are familiar with 80% of the companies listed and possibly even a trusted consumer of their services/products. This will give us some relief and trust when doing financial transactions.

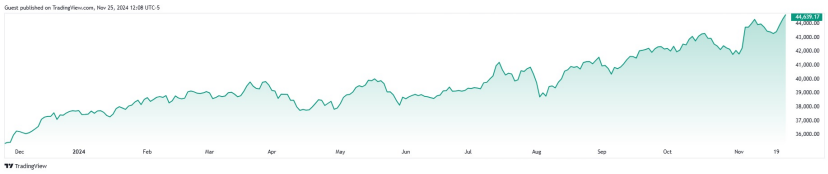

Figure 2.3: TradingView Dow Jones Industrial last year. *n. Select "1 year" tab at the website for current data.*

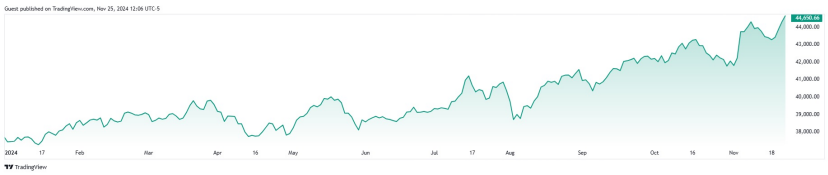

Figure 2.4: TradingView Dow Jones Industrial last year. *n. Select "Year to date" tab at the website for current data.*

One of the most important factors in this list is that the sum of the weights needs to be 1 and so our tests will build portfolios whose weights sum 1. Another important aspect when using these indexes is that the companies included can change at any time so it will be important to review the data online to adjust if required.

We will also include the Dow Jones Titans 50 [16] tickers in our data. As in DJIA, the companies and weights for this index can be changing frequently and we will still be familiar with 70% of them. See the table 2.3 as compiled information from Wikipedia and use the last-mentioned reference [17] for a most updated list.

In summary, We have selected Dow Jones as it has:

- a significantly longer track record (since 1896) even from S&P 500 (since 1957),
- large, well-established, highly reputable companies which demonstrate sustained growth and revenue,
- companies leaders in their respective industries yet small composition, easy to understand and interpret,

[16] Wikipedia. Dow jones global titans 50, 2024b. URL https://en.wikipedia.org/wiki/Dow_Jones_Global_Titans_50. [Online; accessed 30-May-2024]

[17] S&P Dow Jones Indices. Dow jones global titans 50 index factsheet, n.d. URL https://www.spglobal.com/spdji/en/idsenhancedfactsheet/file.pdf?calcFrequency=M&force_download=true&hostIdentifier=48190c8c-42c4-46af-8d1a-0cd5db894797&indexId=1301588. [Online; accessed 30-May-2024]

Table 2.2: Dow Jones Industrial tickers and weights (March 2024)

Company	Ticker	Index weight
3M	MMM	0.0154
American Express	AXP	0.0364
Amgen	AMGN	0.0480
Amazon	AMZN	0.0293
Apple	AAPL	0.0304
Boeing	BA	0.0336
Caterpillar	CAT	0.0545
Chevron	CVX	0.0259
Cisco	CSCO	0.0081
Coca-Cola	KO	0.0102
Disney	DIS	0.0181
Dow	DOW	0.0094
Goldman Sachs	GS	0.0654
Home Depot	HD	0.0623
Honeywell	HON	0.0334
IBM	IBM	0.0309
Intel	INTC	0.0072
Johnson & Johnson	JNJ	0.0270
JPMorgan Chase	JPM	0.0307
McDonald's	MCD	0.0498
Merck	MRK	0.0216
Microsoft	MSFT	0.0683
Nike	NKE	0.0175
Procter & Gamble	PG	0.0269
Salesforce	CRM	0.0504
Travelers	TRV	0.0369
UnitedHealth Group	UNH	0.0881
Verizon	VZ	0.0067
Visa	V	0.0476
Walmart	WMT	0.0100

- companies that must maintain a minimum daily trading volume of 100,000 shares and have been traded on the Nasdaq for at least 2 years

Each index has its strengths and is best suited for different purposes. Experienced investors can modify the extraction parameters for any other index or all tickers as future work.

2.2.2 *Source*

The data source we will use for this research is extracted from TradingView using the Alpaca API using Python. The API collects information from all tickers in the financial market and although there seems to be no limit to the amount of historical data we could extract and the fact we have been able to extract 2 years of data, the amount of data collected and the time to collect it significantly increase as we show in the table 2.4 below:

We will extract 60 to 120 days of Dow Jones tickers data with a 15-

Table 2.3: Dow Jones Industrial Titans 50 tickers (March 2024)

Company	Ticker	Company	Ticker
3M	MMM	JPMorgan Chase	JPM
AbbVie Inc.	ABBV	Mastercard	MA
Allianz	ALV	McDonald's	MCD
Alphabet	GOOG	Merck & Co.	MRK
Amazon	AMZN	Microsoft	MSFT
Amgen	AMGN	Nestlé	NSRGF
Anheuser-Busch InBev	BUD	Novartis	NVS
Apple	AAPL	Nvidia	NVDA
BHP	BHP	Oracle Corporation	ORCL
The Boeing Company	BA	PepsiCo	PEP
BP	BP	Pfizer	PFE
British American Tobacco	BTI	Philip Morris International	PM
Chevron Corporation	CVX	Procter & Gamble	PG
Cisco Systems	CSCO	Roche	ROG
Citigroup	C	Royal Bank of Canada	RY
Coca-Cola	KO	Royal Dutch Shell	SHEL
DuPont	DD	Samsung Electronics	SSUN
ExxonMobil	XOM	Sanofi	SAN
Meta	META	Siemens	ENR
General Electric	GE	Taiwan Semiconductor	TSM
GlaxoSmithKline	GSK	TotalEnergies	TTE
HSBC	HSBC	Visa	V
Intel	INTC	Toyota	TM
IBM	IBM	Walmart	WMT
Johnson & Johnson	JNJ	The Walt Disney Company	DIS

Table 2.4: Alpaca parameters, volume and time results.

# historical days	Frequency	# rows x # columns	Time
120 days	15 Minutes	225,319 rows x 8 columns	0 m 35 s
240 days	15 Minutes	435,565 rows x 8 columns	1 m 25 s
120 days	5 Minutes	582,590 rows x 8 columns	2 m 06 s

minute frequency of updates for these tests as we want the optimization processing time to be low. The more data to process will increase the processing time to optimize and execute the financial transaction which can lead to potential losses for the variable scenario. However, we could collect the data from any or all tickers in the market and for as long as the ticker generated data.

If we would run this in real-time, we would need to consider preloading the historical data preferably segmented by year ranges as there is a limitation on the number of requests we can make over a certain period using the API to prevent abuse and ensure fair access to all users.

It is important to consider each ticker has its own behaviour and historical data so in order for this hypothesis to work with other portfolios or tickers we need to consider that the amount of information for each ticker should stay balanced so we can prevent unnecessary

bias when running the optimization process.

The frequency of updates could be also modified based on the tests we need to run considering the lower frequency will exponentially impact the number of rows collected. We will be using 15 minutes, 30 minutes, integer hours and integer days for our tests.

The data collected will give us 9 different values for each 15-minute frequency chosen from which we will be describing the definition of each in table 2.5 and which ones we will choose for the tests:

timestamp	close	high	low	trade_count	open	volume	vwap	symbol
2022-03-30 08:30:00+00:00	$177.86	$ 177.86	177.82	138	178	2,582	177.84	AAPL
2022-03-30 14:15:00+00:00	$178.91	$ 179.53	178.82	32,350	179	4,045,802	179.11	AAPL
2024-03-28 19:45:00+00:00	$171.48	$ 172.19	171.25	48,487	172	5,115,121	171.80	AAPL
2024-03-28 19:45:00+00:00	$284.27	$ 284.99	284.21	6,430	285	249,972	284.55	AMGN
2024-03-28 19:45:00+00:00	$180.25	$ 180.69	180.16	40,012	181	3,586,599	180.42	AMZN
2024-03-28 19:45:00+00:00	$227.69	$ 228.13	227.66	6,257	228	291,248	227.92	AXP
2024-03-28 19:45:00+00:00	$193.04	$ 193.32	192.84	11,067	193	580,388	193.13	BA
2024-03-28 19:45:00+00:00	$366.27	$ 367.22	366.125	7,211	367	261,950	366.74	CAT
2024-03-28 19:45:00+00:00	$300.99	$ 302.10	300.57	11,173	301	663,371	301.30	CRM
2024-03-28 19:45:00+00:00	$ 49.88	$ 49.98	49.84	17,407	50	2,935,794	49.91	CSCO
2024-03-28 19:45:00+00:00	$157.71	$ 158.09	157.67	11,504	158	847,801	157.92	CVX
2024-03-28 19:45:00+00:00	$122.34	$ 122.55	122.29	14,188	122	1,398,028	122.43	DIS
2024-03-28 19:45:00+00:00	$ 57.93	$ 58.06	57.92	5,655	58	577,929	57.97	DOW
2024-03-28 19:45:00+00:00	$417.57	$ 418.67	417.24	8,027	418	355,038	418.02	GS
2024-03-28 19:45:00+00:00	$383.38	$ 384.19	383.3	9,745	384	497,449	383.75	HD
2024-03-28 19:45:00+00:00	$205.22	$ 205.60	205.18	8,394	205	441,294	205.33	HON
2024-03-28 19:45:00+00:00	$190.90	$ 191.21	190.78	7,745	191	505,045	190.99	IBM
2024-03-28 19:45:00+00:00	$ 44.14	$ 44.35	43.995	46,179	44	8,028,504	44.21	INTC
2024-03-28 19:45:00+00:00	$158.18	$ 158.54	158.15	10,096	159	679,035	158.33	JNJ
2024-03-28 19:45:00+00:00	$200.28	$ 200.72	200.28	14,719	200	847,408	200.52	JPM
2024-03-28 19:45:00+00:00	$ 61.15	$ 61.25	61.11	11,125	61	2,122,107	61.19	KO
2024-03-28 19:45:00+00:00	$281.81	$ 282.20	281.51	8,200	282	526,004	281.92	MCD
2024-03-28 19:45:00+00:00	$106.04	$ 106.20	105.95	8,426	106	605,974	106.08	MMM
2024-03-28 19:45:00+00:00	$131.93	$ 132.18	131.87	11,903	132	1,105,511	132.00	MRK
2024-03-28 19:45:00+00:00	$420.03	$ 421.87	419.92	34,252	421	1,927,758	421.13	MSFT
2024-03-28 19:45:00+00:00	$ 93.97	$ 94.20	93.935	12,482	94	994,116	94.08	NKE
2024-03-28 19:45:00+00:00	$162.16	$ 162.47	162.07	9,377	162	742,311	162.27	PG
2024-03-28 19:45:00+00:00	$230.05	$ 230.87	230.03	3,184	230	142,273	230.52	TRV
2024-03-28 19:45:00+00:00	$494.43	$ 495.28	492.81	10,382	494	477,748	494.23	UNH
2024-03-28 19:45:00+00:00	$ 41.95	$ 42.08	41.95	10,169	42	2,040,959	42.01	VZ
2024-03-28 19:45:00+00:00	$ 278.74	$ 279.80	278.69	9,820	279	545,666	279.42	V
2024-03-28 19:45:00+00:00	$ 60.16	$ 60.34	60.16	16,685	60	2,540,515	60.24	WMT

Table 2.5: Alpaca source data extract sample. *n. Source data varies based on dates chosen.*

`Timestamp` Specific dates and times within the 15-minute frequency we requested are displayed in the UTC zone.

`Close` The closing price is considered significant because it represents the final price agreed upon by buyers and sellers for that trading day. It's often used as a reference point for multiple analyses, charting, trends, return and assessing investment performance.

`High` The highest price at which the ticker was traded during a particular trading session. It reflects the peak price reached during that period.

`Low` The lowest price at which the ticker was traded during a particular trading session. It represents the lowest price level reached during that period.

`Trade_count` The number of trades that occurred for that ticker during a specific trading session. It provides insight into the level of trading activity or liquidity for that transaction.

`Open` The opening price is the price at which it first trades at the beginning of a specific trading session, typically at the start of the trading day. It's another important reference point for analyzing price movements and market sentiment. Similarly to Close, it's often used as a reference point for multiple analysis, charting, trends, return and assessing investment performance.

`Volume` The total number of shares or contracts traded for that ticker during a specific trading session. It indicates the level of trading activity or liquidity in the market for that transaction. We can see the different volume while the market is opened at 14:15 or 19:45 versus 08:30 in table 2.5.

`Wrap` The weighted average price is calculated by multiplying the volume of each trade by its corresponding price, summing up these values, and dividing by the total volume traded during a specific period. It provides a measure of the average price at which all trades occurred, weighted by their respective volumes.

`Symbol` The ticker symbol represents the trading symbol which is a unique combination of letters (and sometimes numbers) assigned to a particular publicly traded company. The length of the ticker symbol is limited so while it is often derived from the company's name it may not always directly match the company's name. Importantly, each ticker symbol is unique within a particular market or exchange. This uniqueness allows investors, traders, and financial institutions to accurately identify and trade the desired ticker.

We will be using the timestamp, close, open and symbol for our tests. We will be creating 2 sets of data one for the open prices and the other for the closed prices leaving each symbol as a column so we can easily access the data.

2.2.3 *Curation*

We are looking for a balance between quality and quantity to ensure valid and generalizable results when doing data curation. We are

looking for data representativeness to preserve the data patterns, and a good and distributed set of data to generalize while still reliable.

The first step is to review whether we have enough information for each of the tickers for which we chose 70% or more as minimum information to keep the ticker. It would be highly impossible to find any of the Dow Jones tickers with less than 70% information and we would need to question the source if that would happen. The 60% is used as the minimum percentage recommended for any analysis to guarantee the quality of our results. All the tickers used for these tests confirmed more than 92%; however, we decided to keep 70% as a minimum in the code for any other tickers used in future work to guarantee the balance between quality and quantity.

The next step is (through observation and later confirmed), to only keep the timestamps within the active trading (high volume) which were between Monday to Friday and from 14:30 to 21:00 UTC. Using the NYSE as a reference[18], we confirmed the active market was open from 9:30 am to 4:00 pm Eastern Standard Time from Monday to Friday. We have also noticed that for some of the tickers, the amount of data collected after 20:00 UTC was noticeably reduced for which we adjusted it while we looked for more reference information that can explain this behavior.

[18] NYSE. Hours & calendars, 2024. URL https://www.nyse.com/markets/hours-calendars. [Online; accessed 30-May-2024]

On the other hand, we also need to account for the NYSE holidays[19] posted on its website as there would be no data available during those days and none of the scenarios nor any transactions could be executed during those dates, for which we will be using the next date available.

[19] NYSE. Hours & calendars, 2024. URL https://www.nyse.com/markets/hours-calendars. [Online; accessed 30-May-2024]

We see transactions happening out of those timestamps[20]; however, for this research, we decided any of the transactions would be during the active market timestamp for which we only kept data within this date range to keep the data behaviour balanced.

[20] Investopedia. Trading session, 2024. URL https://www.investopedia.com/terms/t/tradingsession.asp. [Online; accessed 30-May-2024]

2.2.4 *Quality*

The next step is to analyze the data and identify if there could be any problems and how we will mitigate or prevent the problem. We run a data quality report for each of the data sets we are using, see the results of one sample in table 2.6:

Missing values (Nan value) Any number above 0 in this column represents a missing price for a specific timestamp for that ticker. We decided to always take the previous price available for any missing price as it is also a common practice when trading.

Name	Type	Nan value	Count values	Unique values	Min	Max	Mean	Std	Var	Median	Skewness	Kurtosis	IQR Min	IQR Max	IQR Num
MMM	float64	0	12,810	6,843	85.49	154.34	113.73	17.82	317.59	108.12	0.67	-0.62			
AXP	float64	311	12,499	6,773	132.78	231.49	165.76	20.31	412.60	160.67	1.23	1.22			
AMGN	float64	393	12,417	7,760	212.89	327.74	256.61	22.69	514.65	252.32	0.51	-0.26			
AMZN	float64	2	12,808	9,351	81.56	3,366.93	345.68	723.89	524,012.12	128.05	3.07	7.82			
AAPL	float64	0	12,810	8,236	124.29	199.38	165.39	18.78	352.87	167.81	-0.14	-1.12			
BA	float64	51	12,759	9,054	113.30	265.95	190.24	32.60	1,062.80	199.99	-0.34	-0.45			
CAT	float64	318	12,492	9,130	160.74	366.77	241.39	42.72	1,824.73	237.33	0.56	0.04			
CVX	float64	102	12,708	5,805	133.08	189.37	160.55	11.23	126.01	159.64	0.23	-0.60			
CSCO	float64	116	12,694	4,037	38.96	58.10	48.85	3.68	13.51	49.10	-0.15	-0.21			
KO	float64	91	12,719	3,345	51.67	66.93	60.67	2.63	6.92	60.55	-0.51	0.26			
DIS	float64	11	12,799	6,634	78.83	142.73	98.48	11.61	134.70	96.15	0.87	0.79			
DOW	float64	304	12,506	4,010	43.09	71.39	54.15	5.26	27.67	53.33	1.05	1.29			
GS	float64	218	12,592	8,306	278.52	418.11	340.35	28.79	828.77	335.96	0.33	-0.68			
HD	float64	296	12,514	7,980	265.87	396.79	312.14	26.68	711.57	306.37	0.90	0.29			
HON	float64	375	12,435	5,762	166.97	220.96	195.24	10.71	114.76	195.62	-0.15	-0.20			
IBM	float64	241	12,569	6,157	117.12	198.76	143.16	18.07	326.50	138.62	1.42	1.37			
INTC	float64	2	12,808	5,407	24.77	52.36	35.64	6.86	47.12	34.61	0.36	-1.06			
JNJ	float64	240	12,570	5,010	145.13	186.45	165.01	9.05	81.91	163.59	0.21	-0.97			
JPM	float64	78	12,732	7,215	101.96	200.67	140.41	20.42	416.99	138.50	0.72	0.20			
MCD	float64	322	12,488	6,839	228.85	301.90	271.62	17.61	310.24	271.21	-0.24	-1.02			
MRK	float64	273	12,537	5,748	81.84	132.24	104.64	11.50	132.28	106.43	-0.05	-0.67			
MSFT	float64	1	12,809	9,646	214.25	430.16	304.62	56.05	3,141.35	288.58	0.50	-0.89			
NKE	float64	187	12,623	5,728	82.85	139.45	109.91	10.74	115.34	108.17	0.16	-0.46			
PG	float64	288	12,522	4,823	122.44	164.77	147.98	7.91	62.63	148.91	-0.50	0.15			
CRM	float64	166	12,644	9,110	126.68	318.38	199.59	45.21	2,043.76	194.44	0.75	-0.10			
TRV	float64	528	12,282	5,887	149.82	230.60	179.05	16.46	271.00	175.37	1.15	1.02			
UNH	float64	308	12,502	8,070	447.68	557.60	508.32	23.85	569.00	508.52	-0.08	-1.01			
VZ	float64	43	12,767	5,132	30.38	55.43	39.82	5.41	29.29	38.66	0.93	0.24			
V	float64	305	12,505	7,803	177.18	290.84	228.52	25.36	643.35	226.25	0.48	-0.32			
WMT	float64	192	12,618	6,409	58.25	180.23	143.55	22.79	519.33	149.16	-2.32	6.15			

Table 2.6: Data quality report. *n. Source data varies based on dates chosen.*

`Count values` We have enough and on average a similar amount of data for each ticker which will guarantee a balanced exercise when running any process for the data. If any of the tickers would have a significantly lower or higher number, we would need to review the specific situation for that ticker and how it could affect the results.

`Unique values` We can see good price fluctuation which will guarantee different top tickers when running the exercise with different time frequency. If any of the tickers would have a lower number in this column, we would need to dismiss the ticker as it wouldn't provide any relevant information to the results.

`Measures of position, central tendency and dispersion` These measures of position (Min and Max), measures of central tendency (Mean and Median) and measures of dispersion (Std and Var) of the prices can show us how the price has fluctuated over time; and as such the standard deviation and variance will be used to calculate the volatility of the ticker as well as the calculated weights for the optimization process. The correlation between these measures refers to some degree to their risks and how their behaviors in the market could be connected. If you want to analyze further any of these numbers, you would need to evaluate the changes in the market, emotions and risks for each of the tickers that have more variance.

`Skewness and kurtosis` The statistical measures help us describe the shape and distribution of the prices and the "ideal" values in these columns vary on the context of the problem we need to resolve. Since we are not going to try to predict a price in this research, none of these columns impact the results of what we will be running. We

have highlighted in light blue 3 top tickers we would need to focus on for further analysis in case we were focusing on predicting prices as part of the data curation. As we can observe, both their skewness and kurtosis are very different than the rest of the tickers so those could impact the results.

Outliers This last quality number confirms our data is ready to go as we don't see any outliers (minimum, maximum and count) so no further data transformation is required. If any of the prices would be shown as outliers, we would have needed to understand how including or excluding this/these price(s) would affect our results.

3 Optimization

Contents

The Markowitz Optimization Theory, also known as Modern Portfolio Theory (MPT) transformed the investment strategies using a mathematical optimization method which was introduced by Harry Markowitz in 1952. [1] It considers as primary values the returns of the tickers, and the correlation between their open or close prices so as a result it can provide the optimal tickers and weights to obtain the highest return at the lowest risk for the given returns. By investing in more than one stock, an investor can reap the benefits of diversification, and reduce the risk of the portfolio [2].

[1] H. Markowitz. Portfolio selection, 1952. URL https://www.jstor.org/stable/2975974. [Online; accessed September 8, 2024]

[2] Investopedia. Investopedia, Year the website was last updated. URL https://www.investopedia.com/. Accessed on Date

The expected return of the portfolio is calculated as a weighted sum of the individual assets' returns, while the portfolio's risk is measured by the variance and covariances of the asset returns and represented generally and mathematically as follows:

$$\min_{\mathbf{w}} \mathbf{w}^\top \boldsymbol{\Sigma} \mathbf{w} \tag{3.1}$$

Subject to the following constraints:

$$\begin{aligned} \mathbf{w}^\top \boldsymbol{\mu} &\geq R_t, \\ \sum_i w_i &= 1, \\ w_i &\geq 0. \end{aligned}$$

Where:

- $\mathbf{w}$ is the vector of portfolio weights
- $\boldsymbol{\Sigma}$ is the covariance matrix of asset returns
- $\boldsymbol{\mu}$ is the vector of expected returns for each asset
- R_t is the target portfolio return
- w_i is the capital (investment)
- $\mathbf{w}^\top \boldsymbol{\mu}$ is the portfolio return
- $\mathbf{w}^\top \boldsymbol{\Sigma} \mathbf{w}$ is the portfolio variance (risk)

And the objective function is to minimize the portfolio variance ($\mathbf{w}^\top \boldsymbol{\Sigma} \mathbf{w}$) while meeting or exceeding the target return(R_t).

MPT comes with some detractors as in the real financial life the behaviour of the stock value will not have a normal distribution and the risk is an unpredictable market events and behaviors. As this research is mainly focused on non-experienced investors, we will assume the calculated volatility represents "accurately" our financial risk (although we know that risk in the financial markets is a multivariable factor).

As in any optimization mathematical process, we should include our constraints to the objective function such as the investment available (cannot be exceeded), and the number of tickers to consider (cannot be exceeded, it can be reduced instead as per the budget).

We could have used the Monte Carlo simulation method to generate a wide number of possible outcomes randomly which we could translate into multiple returns and as such select the weight of our stocks based on what we believe could be the closest realistic scenario.

However, we decided to keep it focused on Markowitz equation (3.1) for which we coded all of the steps within Python and described with words it consists on the following steps:

1. Collect historical price data (behavior of the assets in the past),

2. Calculate returns (average price of each ticker) and covariance (relationship between the tickers),
3. Define the variables if applicable like the investment for each stock. In this research, we will use the 100% investment based on the calculated weights the optimization process provides,
4. Define the constraints such as the sum of weights should always be equal to 1 (or 100%), the number of tickers and the investment cannot be exceeded, the financial transaction should happen only during open market time zones and holidays are excluded
5. Define the objective function to maximize the return and minimize the portfolio variance (risk),
6. Apply the optimization process and method of Lagrange multipliers to the convex optimization (minimization)
7. Finally, get the best weights for each stock to maximize our returns.

In the context of MPT, investors aim to construct the optimal portfolio (weights for each stock) that maximizes expected return for a given level of risk. Covariance is used to calculate the overall risk (volatility) of the portfolio, which is not simply the sum of individual risks but also includes the degree to which each asset's returns co-vary with every other asset in the portfolio:

1. Positive Covariance: If two assets have a positive covariance, their returns tend to move in the same direction. If one asset's return goes up, the other is likely to go up as well, and vice versa.
2. Negative Covariance: Conversely, a negative covariance indicates that the returns of two assets move in opposite directions. When the return on one asset goes up, the return on the other tends to go down.
3. Zero Covariance: A zero covariance means that the returns of the two assets show no linear relationship and are independent of each other.

By including assets with low or negative covariance, investors can reduce the portfolio's overall risk (by diversification). This is because when one asset experiences a loss, an asset with a negative covariance might gain, offsetting the loss.

Finally, this model also talks about how the covariance is used to determine the efficient frontier, which represents the set of portfolios

that offer the highest expected return for a given level of risk. Portfolios along the efficient frontier are considered optimal because they are expected to provide the best possible return for the risk taken.

3.1 Variables

Understanding the following 3 concepts is crucial for assessing the potential profitability and risk of an investment for anybody who wants to create portfolios.

3.1.1 Returns

It represents the gain or loss on an investment over a certain period, typically expressed as a percentage between [-100, ∞).

$$Returns = \frac{P_i - P_{i-1}}{P_{i-1}} \tag{3.2}$$

Where:

- P_i is the price at time i.
- P_{i-1} is the price at the previous time period.

3.1.2 Volatility (Risk)

It measures the degree of variation in the returns of an investment over time, which we commonly name risk. High volatility means higher risk while lower volatility suggests stability and directly relates to variance and standard deviation. Values go between [0, ∞).

$$\text{Volatility} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(r_i - \bar{r})^2} \tag{3.3}$$

Where:

- n is the number of observations.
- r_i is the return at time i.
- $\bar{r}$ is the average return over the period.

3.1.3 Sharpe ratio

Sharpe ratio[3] is a measure of risk-adjusted return, which assesses the return of an investment relative to its risk. Values go between $[0, \infty)$.

$$\text{Sharpe Ratio} = \frac{R_p - R_f}{\sigma_p} \tag{3.4}$$

Where:

- R_p is the average return of the portfolio.
- R_f is the risk-free rate of return.
- σ_p is the standard deviation of the portfolio's excess return.

A risk-free rate represents the interest an investor would expect from a risk-free investment over a specified period. In practice, the risk-free rate of return does not truly exist so we can subtract the inflation rate from the ticker over the duration.

Considering inflation introduces normally uncertainty into the markets that will be automatically reflected in the stock price (commonly as an inverse correlation) and the tests for this research will be over small periods going from hours to maximum days, so our Sharpe ratio will be simplified as returns divided by the standard deviation of the portfolio; considering risk-free rate as 0.

[3] W. F. Sharpe. A theory of market equilibrium under conditions of risk, 1964. URL https://www.jstor.org/stable/2977928. [Online; accessed September 8, 2024]

3.2 Adjustments

In order for the optimization to run efficiently and effective, we also need to consider certain adjustments to simplify, limit and/or guarantee the optimization will behave evenly regardless of the conditions.

3.2.1 Stocks

Many documents are talking about the "right" number of stocks a portfolio should have and it will be hard to generalize it. The perfect number will always depend on the investor's situation and needs for the investment, amount of time to invest, expectation of returns and/or accepted risk.

The more stocks we have in the portfolio will certainly reduce the risk; however, we may need to pay more fees and similarly introduce another complexity we would also advise buying a full stock rather

than partial stock as depending on the stock and broker, some will allow to buy partial stocks while others won't allow it or may charge more to do that.

We have simplified this research to buy only full stocks and keep between 15 to 30 stocks in our selected portfolio [4]. And we will notice the portfolio will reduce due to the restriction of only buying full stocks.

[4] Investopedia. Optimal portfolio size, n.d. URL https://www.investopedia.com/ask/answers/05/optimalportfoliosize.asp. [Online; accessed June 7, 2024]

3.2.2 *Investment*

Following the path for simplification and considering all data source data is represented in United States dollars (USD), our investment parameter will be always in USD and we advise this to stay above $ 100,000 USD. A lower investment could reduce the portfolio when buying only full stocks and generate different results.

We will be comparing the results of 2 scenarios for which we will have an initial investment and a defined period, we will maximize the investment to trade full stock and capture the results by subtracting the final sell investment minus the initial investment and adding the remainder of the investment.

As an example, see table 3.1 below with an initial investment of $ 10,000 USD and we have chosen to only use 6 stocks for our portfolio, let's analyze how the test will run with these parameters:

Initial investment $ 10,000.00

Ticker	Weight	Open price	estment per weil	previous remin	Sub-total	# of stocks	Remainder	Actual weight
MSFT	0.32	$ 421.37	$ 3,200.00	$ -	$ 3,200.00	7	$ 250.41	0.29
HON	0.19	$ 205.18	$ 1,900.00	$ 250.41	$ 2,150.41	10	$ 98.61	0.21
IBM	0.18	$ 191.20	$ 1,800.00	$ 98.61	$ 1,898.61	9	$ 177.81	0.17
PG	0.13	$ 162.30	$ 1,300.00	$ 177.81	$ 1,477.81	9	$ 17.15	0.15
AAPL	0.10	$ 172.10	$ 1,000.00	$ 17.15	$ 1,017.16	5	$ 156.68	0.09
NKE	0.06	$ 93.97	$ 600.00	$ 156.68	$ 756.68	8	$ 4.96	0.08
CRM	0.02	$ 301.24	$ 200.00	$ 4.96	$ 204.96	0	$ 204.96	0.00

Table 3.1: Investment optimization process. *n. Try using all investment to the maximum.*

`Weight` Our optimization process will give us the suggested percentage we should invest into that ticker so 32% of our investment should be used for Microsoft (MSFT) of the $10,000.

`Open price` This is the price our source data will give us for that specific time we are running our test.

Investment per day This is the multiplication of the 0.32 weight and the $ 10,000 USD investment so we know how much money we should use to buy from this stock.

Add previous reminder When buying the very first stock, we won't have any reminder for which we are representing in $ 0 USD. Let's review what happened after buying the first stock; we have $ 3,200 USD to buy stocks from Microsoft with a price of $ 421.37 USD per stock and considering we can only buy full stocks; we can only buy 7 stocks (7 * $ 421.37 = $ 2,949.59) so we would have a reminder of $ 250.41 USD. These $250.41 will be added to the next trade for Honeywell (HON), we originally had $ 1,900 USD which would allow us to buy only 9 stocks; however, we will add the previous reminder and now we have $ 2,150.41 USD to buy 10 stocks.

Sub-total It represents the new investment to buy stocks which is the sum of the original investment per weight plus the previous reminder.

# of stocks Full stocks are bought with the new investment for that stock.

Actual weight Net weight for each stock in the portfolio considering the full stock trade and we can see the Salesforce stock (CRM) has been discarded as we only have $ 200.00 plus the $ 4.96 USD reminder (from previous stock) and the CRM stock price is $ 301.24 USD so we cannot buy anymore.

As we can see in the example above, we will finalize our full trade with only 6 stocks as we won't have enough new investment to buy a Salesforce (CRM) stock for $ 301.24 USD and we will keep a reminder of $ 204.96 USD we won't be able to invest.

For this research, we will give priority to the weight and we will leave the final reminder without investing even if there could be a way to still buy more stocks as we want to keep as many tickers as we can as close as possible to the original weight.

One last important point, since the initial investment for the "static" scenario will remain invested until the end of the period we chose; our partial new investments for the "variable" scenario will be fully invested too which means any gains or losses will be reinvested plus the reminders.

As a possible adjustment, we could introduce a different mechanism to use the reminders of $ 204.96 USD left; we could decide to buy one

more stock from either IBM, PG, AAPL or NKE. We can then compare the initial weight with the actual weight and discard PG and NKE as their actual weights are already higher than their suggested initial weights. Further tests would need to be evaluated to incorporate this into our model and possible associated risks.

There are multiple ways to use the reminders based on the priorities we want to incorporate: weight, number of tickers in the portfolio or higher number of stocks. Each of them will possible impact the results and it will be important to run both static and variable scenarios and then compare the results.

3.2.3 *Price*

The choice between using open or close price when trading by minutes, hours or days depends on the trading strategy and objectives and both are commonly discussed in trading and financial markets.

As per several articles and literature, the close price is normally used to assess where the market settled and assess the day's performance to plan for the next day while the open price is normally used to identify initial market sentiment and potential opportunities or direction.

We will be using a "close" price for our "static" scenarios and an "open" price for our "variable" scenarios. However, the process will provide further results for us to observe and adjust if necessary considering all our tests will be using minutes, hours and days.

It is also important to mention that for those scenarios our specific price for that particular period is unavailable, we will be using the previous price available as a common practice in the trading market.

3.2.4 *Valid periods*

Although the trading market and some brokers allow investors to buy or sell after hours, we will be exclusively running our scenarios during active market days and hours.

So, every start and end date time provided will be translated into a valid date and time from an active market period.

In the table below 3.2, the start date time was March 16 at 8:30 AM which is a Saturday, so the next valid active market time is March 18 at 2:30 PM and this becomes our new initial start date time.

Considering our frequency is every 3 hours, we will add the 3 hours to our new initial start date time and transform that next date time into a valid date time until we reach our end date time.

As you can see on the left side of the table there were 4 invalid date-time which would impact our results so we converted these parameters into what you can see on the right side of the table with valid date-time periods.

Start	Mar 16, 2024 08:30
End	Mar 20, 2024 23:30
Frequency	3 hours

Date	Day	Note	Next valid date and time	Day
Mar 16, 2024 08:30	Saturday	Invalid date and time	Mar 18, 2024 14:30	Monday
Mar 18, 2024 17:30	Monday	Valid	Mar 18, 2024 17:30	Monday
Mar 18, 2024 20:30	Monday	Valid	Mar 18, 2024 20:30	Monday
Mar 18, 2024 23:30	Monday	Invalid time	Mar 19, 2024 14:30	Tuesday
Mar 19, 2024 17:30	Tuesday	Valid	Mar 19, 2024 17:30	Tuesday
Mar 19, 2024 20:30	Tuesday	Valid	Mar 19, 2024 20:30	Tuesday
Mar 19, 2024 23:30	Tuesday	Invalid time	Mar 20, 2024 14:30	Wednesday
Mar 20, 2024 17:30	Wednesday	Valid	Mar 20, 2024 17:30	Wednesday
Mar 20, 2024 20:30	Wednesday	Valid	Mar 20, 2024 20:30	Wednesday
Mar 20, 2024 23:30	Wednesday	Invalid time	Mar 21, 2024 14:30	Thursday

Open market days	Monday to Friday
Open market hours	14:30 to 21:00

Valid date time table		
Start datetime	End datetime	# of trading
Mar 18, 2024 14:30	Mar 18, 2024 17:30	1
Mar 18, 2024 17:30	Mar 18, 2024 20:30	2
Mar 18, 2024 20:30	Mar 19, 2024 14:30	3
Mar 19, 2024 14:30	Mar 19, 2024 17:30	4
Mar 19, 2024 17:30	Mar 19, 2024 20:30	5
Mar 19, 2024 20:30	Mar 20, 2024 14:30	6
Mar 20, 2024 14:30	Mar 20, 2024 17:30	7
Mar 20, 2024 17:30	Mar 20, 2024 20:30	8
Mar 20, 2024 20:30	Mar 21, 2024 14:30	9

Table 3.2: Valid periods calculation. *n. Guarantee only valid periods are used.*

We have now our valid date timetable and we have marked in light blue the start and end date time for our "static" scenario and our 9 number of trading start and end date time ranges for our "variable" scenario.

3.3 *Fees and costs*

As per Cagan, Michele [5][p. 186], the market needed an association that could regulate the exchange as well as certify the financial professionals doing that so FINRA was created in July 2007 when the National Association of Securities Dealers (NASD) merged with the regulation committee of the NYSE.

We have specifically considered platforms, data, optimization process and adjustments to generate zero or minimum fees and/or costs: Alpaca as a brokerage doesn't charge for deposits or withdrawals and the only fees applicable will be linked specifically to these regulatory entities SEC and FINRA and none of the Dow Jones Industrial stocks will charge any of the fees for trading stocks.

There are also some fees or tax calculators available for which we have confirmed none of the Dow Jones will generate any extra fees or taxes; although we could get 100% certainty once we run this in real-time using Alpaca [6].

[5] M. Cagan. *Stock Market 101: From Bull and Bear Markets to Dividends, Shares, and Margins—Your Essential Guide to the Stock Market*. Adams Media, kindle edition edition, 2016

[6] BrokerChooser. Brokerage fee calculator, n.d. URL https://brokerchooser.com/brokerage-fee-calculator. [Online; accessed June 7, 2024]

4 Results & Code

Contents

Let's remember our hypothesis,

- non-experienced investors,
- using optimization process,
- selecting between 15 to 30 tickers from the Dow Jones market index,
- buying only full stocks,
- during the valid open day and time market,
- within a period

could get higher net gains or lower net losses by trading multiple times (variable scenario) than only trading one time (static scenario).

We also mentioned that based on common practice the "static" scenario will use the close price while the "variable" scenario the open price as that is the suggested way based on common practice.

We used one set of parameters between March and May 2024 to demonstrate the hypothesis between static and variable scenarios. Those preliminary results helped us to provide some assumptions and gave us enough data to continue running other sets of parameters in future months. The next sets of parameters between September and November 2024 confirmed the initial assumptions were still valid based on the new results captured.

4.1 Preliminary results

As you can see in figure 4.1, after running 40 scenarios with multiple parameters, 34 of 40 (85.00%) of the cases were successful for which the sum of the variable scenarios gained $ 239,611.41 USD, $ 64,944.02 USD above the total gain from the static scenarios.

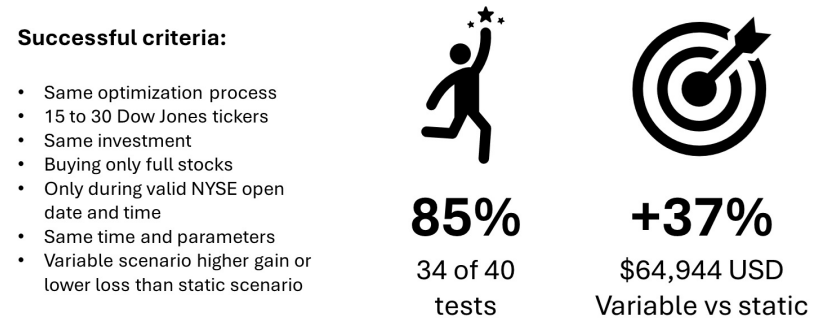

Figure 4.1: Summary of the preliminary results. *n. All variables and adjustments included.*

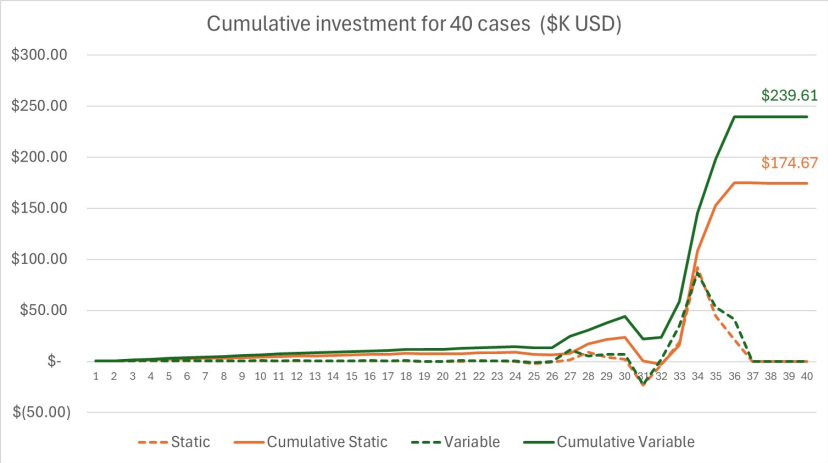

Figure 4.2: Cumulative investment for the 40 tests. *n. All variables and adjustments included.*

By examining these 40 cases in detail, we can generate an initial optimal portfolio to purchase stocks. For the static scenarios, these 40 cases involve selling the stocks at the end date and subtracting the initial investment to determine the net gain or loss. In contrast, for the variable scenarios, these 40 cases involve selling the stocks, generating a new optimal portfolio, and purchasing new stocks multiple times until the final sell at the end date, then subtracting the initial investment to determine the net gain or loss. Ultimately, we sum the total net gain

or loss from these 40 cases, resulting in $ 174,667.26 USD for the static scenario and $ 239,611.41 USD for the variable scenario as we can see in figure 4.2.

Due to the limitation of only buying full stocks, we noticed the maximum number of tickers for the variable scenario was 23 while for the static scenario was 20 regardless of using the same parameters.

Scenario	Start datetime	End datetime	Step	Step type	Tickers	Investment
0	3/4/2024 14:30	3/4/2024 17:30	15	Minutes	25	$ 100,000.00
1	3/4/2024 14:30	3/4/2024 17:30	30	Minutes	25	$ 100,000.00
2	3/4/2024 14:30	3/4/2024 17:30	1	Hours	25	$ 100,000.00
3	3/4/2024 14:30	3/6/2024 17:30	3	Hours	25	$ 100,000.00
4	3/4/2024 14:30	3/8/2024 17:30	1	Days	25	$ 100,000.00
5	3/4/2024 14:30	3/11/2024 17:30	4	Hours	25	$ 100,000.00
6	3/4/2024 14:30	3/4/2024 17:30	15	Minutes	30	$ 100,000.00
7	3/4/2024 14:30	3/4/2024 17:30	30	Minutes	30	$ 100,000.00
8	3/4/2024 14:30	3/4/2024 17:30	1	Hours	30	$ 100,000.00
9	3/4/2024 14:30	3/6/2024 17:30	3	Hours	30	$ 100,000.00
10	3/4/2024 14:30	3/8/2024 17:30	1	Days	30	$ 100,000.00
11	3/4/2024 14:30	3/11/2024 17:30	4	Hours	30	$ 100,000.00
12	3/4/2024 14:30	3/4/2024 17:30	15	Minutes	20	$ 100,000.00
13	3/4/2024 14:30	3/4/2024 17:30	30	Minutes	20	$ 100,000.00
14	3/4/2024 14:30	3/4/2024 17:30	1	Hours	20	$ 100,000.00
15	3/4/2024 14:30	3/6/2024 17:30	3	Hours	20	$ 100,000.00
16	3/4/2024 14:30	3/8/2024 17:30	1	Days	20	$ 100,000.00
17	3/4/2024 14:30	3/11/2024 17:30	4	Hours	20	$ 100,000.00
18	3/14/2024 14:30	3/19/2024 14:30	1	Days	30	$ 100,000.00
19	3/15/2024 14:30	3/18/2024 17:30	4	Hours	30	$ 100,000.00
20	3/15/2024 14:30	3/20/2024 17:30	1	Days	30	$ 100,000.00
21	3/16/2024 14:30	3/21/2024 17:30	1	Days	30	$ 100,000.00
22	3/17/2024 14:30	3/22/2024 17:30	1	Days	30	$ 100,000.00
23	3/18/2024 14:30	3/23/2024 17:30	1	Days	30	$ 100,000.00
24	3/14/2024 14:30	3/19/2024 14:30	1	Days	30	$ 1,000,000.00
25	3/15/2024 14:30	3/18/2024 17:30	4	Hours	30	$ 1,000,000.00
26	3/15/2024 14:30	3/20/2024 17:30	1	Days	30	$ 1,000,000.00
27	3/16/2024 14:30	3/21/2024 17:30	1	Days	30	$ 1,000,000.00
28	3/17/2024 14:30	3/22/2024 17:30	1	Days	30	$ 1,000,000.00
29	3/18/2024 14:30	3/23/2024 17:30	1	Days	30	$ 1,000,000.00
30	3/14/2024 14:30	3/19/2024 14:30	4	Hours	30	$ 10,000,000.00
31	3/15/2024 14:30	3/18/2024 17:30	4	Hours	30	$ 10,000,000.00
32	3/15/2024 14:30	3/20/2024 17:30	4	Hours	30	$ 10,000,000.00
33	3/16/2024 14:30	3/21/2024 17:30	4	Hours	30	$ 10,000,000.00
34	3/17/2024 14:30	3/22/2024 17:30	4	Hours	30	$ 10,000,000.00
35	3/18/2024 14:30	3/23/2024 17:30	4	Hours	30	$ 10,000,000.00
36	4/1/2024 14:30	4/3/2024 14:30	2	Hours	30	$ 10,000.00
37	4/1/2024 15:30	4/3/2024 15:30	2	Hours	30	$ 10,000.00
38	4/1/2024 18:00	4/3/2024 18:00	2	Hours	30	$ 10,000.00
39	4/1/2024 18:00	4/4/2024 14:30	2	Hours	30	$ 10,000.00

Table 4.1: Preliminary parameters for the 40 tests. *n. All variables and adjustments included.*

We also measured the associated time required to process the

optimization, select the tickers and trading process for which we noticed it took 02:14 minutes on average for the variable scenario while it took 00:03 minutes for the static scenario.

We can analyze the parameters for the 40 cases used for the preliminary results, as we can see in table 4.1 each of the parameters were moved to capture changes in behaviors that can help us adjust the parameters better to create greater gains. Each of these parameters should provide enough flexibility to the non-experienced investor as well as for the process to adjust to the market changes.

See the last 4 cases in table 4.2 we run, everything remains the same except for the starting and end times for which we can see the combination of the ending time and either the close or open price can make a difference and will require further analysis.

Static with close price

Scenario	Start datetime	End datetime	Step	Step type	Tickers	Investment	Tickers selected	Actual investment	Remainder	Final sell	Total gain/loss	Time cost
36	4/1/2024 14:30	4/3/2024 14:30	2	Hours	30	$ 10,000.00	17	$ 9,996.81	$ 3.19	$ 9,985.07	$ (8.56)	00:22.8
37	4/1/2024 15:30	4/3/2024 15:30	2	Hours	30	$ 10,000.00	16	$ 9,996.81	$ 3.19	$ 9,977.45	$ (16.17)	00:27.1
38	4/1/2024 18:00	4/3/2024 18:00	2	Hours	30	$ 10,000.00	14	$ 9,966.16	$ 33.84	$ 9,941.98	$ 9.67	00:27.2
39	4/1/2024 18:00	4/4/2024 14:30	2	Hours	30	$ 10,000.00	15	$ 9,999.06	$ 0.94	$ 10,018.65	$ 20.53	00:36.4

Variable with open price

Scenario	Start datetime	End datetime	Step	Step type	Tickers	Investment	Tickers selected	Actual investment	Remainder	Final sell	Total gain/loss	Time cost
36	4/1/2024 14:30	4/3/2024 14:30	2	Hours	30	$ 10,000.00	17	$ 9,996.65	$ 31.02	$ 9,958.57	$ (7.07)	03:57.9
37	4/1/2024 15:30	4/3/2024 15:30	2	Hours	30	$ 10,000.00	14	$ 9,980.64	$ 33.33	$ 9,963.47	$ 16.16	04:09.3
38	4/1/2024 18:00	4/3/2024 18:00	2	Hours	30	$ 10,000.00	13	$ 9,998.96	$ 21.03	$ 9,948.16	$ (29.77)	03:58.7
39	4/1/2024 18:00	4/4/2024 14:30	2	Hours	30	$ 10,000.00	13	$ 9,998.96	$ 20.08	$ 9,990.82	$ 11.95	04:55.6

Table 4.2: 4 case comparison between open and close price. *n. All variables and adjustments included.*

4.2 *Results*

We monitored the market, confirming an increase in the Dow Jones Index since May 2024, which has driven a general rise in stock prices. Due to market uncertainty and rising prices, the number of selected tickers in our process decreased even if we increased the investment up to $ 1,000,000 USD.

We reviewed and optimized the code and conducted 96 tests using October dates and historical data (prices) from September to November 2024. As you can see in table 4.3 there were 62% successful cases for variable scenarios from which 48% were using close price instead of open; however the associated cost for processing time is significantly higher than the open price one.

We are sharing the 4 possible scenarios as the market changed and we mentioned this optimization process should constantly adapt to

Table 4.3: Summary of results in average (AVG)

Scenario	# Win scenarios	# Tickers (AVG)	Minutes (AVG)
`Static with close price`	11	7	00:19
`Static with open price`	23	11	00:20
`Variable with close price`	44	7	02:43
`Variable with open price`	18	7	01:56

new conditions. More importantly, stock prices increased so variance and covariances were impacted as well as our selected portfolio. We see this impact in the number of tickers as even if our scenarios asked for 15, 20, 25 and 30 tickers the combination of the weight & price for each ticker plus the full stock limitation rounded this number down to less than 10 in average for 96 tests. The max number of tickers in static scenarios was 19 and for variable scenarios was 13.

If we focused exclusively on the 2 initial scenarios, the variable scenario with open price got 4% more gains and 54% successful rate than the static scenario with close. Even if the numbers were less favorable than the ones we captured preliminary, they still confirm our hypothesis.

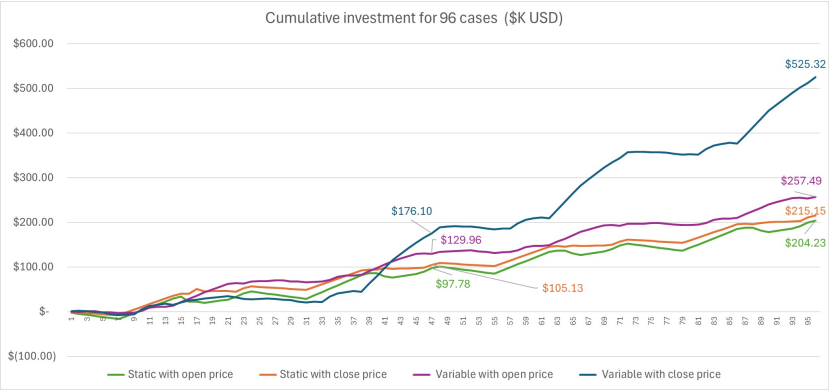

Figure 4.3: Cumulative investment for the final 96 tests. *n. All variables and adjustments included.*

Let's now focus on the cumulative gains and losses in figure 4.3, the 96 cases were purposely divided into 4 sections and the only difference between sections was the number of tickers. Test cases 0 to 23 with 15 tickers, 24 to 47 with 20 tickers, 48 to 71 with 25 tickers and 72 to 95 with 30 tickers. Dividing the chart into four sections helps illustrate the cumulative gains across different tests. In test 23, the gains for both the static and variable scenarios are nearly identical. By test 47, however, both variable scenarios start to diverge from the static scenario, and this gap widens significantly by test 96. The chart also highlights how

the static scenario maintains a stable trend from test 0 to 96, with the number of tickers showing minimal impact on its performance.

However, the variable scenario using either close or open price significantly impacted the cumulative gains and understanding why using the close price significantly generates higher net gains than using the open price requires further analysis.

In summary, the results show the variable scenario generates more net gains than the static scenario and confirm how important the optimization process adjustment to the market is and will continue to be.

4.3 Code & Files

The code and entry files used for this document are listed in the table 4.4 below and are shared in GitHub[1]. Plus we are adding some samples of files the code generates for us to analyze the results from different perspectives: gains and losses, overall results from the parameters perspective and details for each financial transaction required. As we can realize, there is plenty of data for us to continue exploring options to refine and obtain better results and we always have the opportunity to keep improving.

[1] Rosmarsan. Optimized pathway for non-experienced investors, 2024. URL https://github.com/rosmarsan/Optimized-pathway.git. [Online; accessed 30-May-2024]

Table 4.4: List of documents available in GitHub

File name
Optimized_pathway_for_non_experienced_investors_v2.3.pdf
Scenarios.xlsx
NYSE Holidays.xlsx
Data_quality_clean_prices_open.xlsx
Results_test for file Scenarios - Part 1 - 4_18_static_close.xlsx
Results_test for file Scenarios - Part 1 - 4_18_static_open.xlsx
Results_test for file Scenarios - Part 1 - 4_18_variable_close.xlsx
Results_test for file Scenarios - Part 1 - 4_18_variable_open.xlsx
Test results for file Part 2 - 4.xlsx
Test summary for file Scenarios - Part 2 - 4.xlsx

Optimized_pathway_for_non_experienced_investors_v2.3.pdf Code used to run all test scenarios used and referenced in this document.

Scenarios.xlsx Entry file that allows you to test multiple scenarios at a time with different parameters.

NYSE Holidays.xlsx Entry file with the list of holidays published on the NYSE website[2].

[2] NYSE. Hours & calendars, 2024. URL https://www.nyse.com/markets/hours-calendars. [Online; accessed 30-May-2024]

Data_quality_clean_prices_open.xlsx Sample of the data quality analysis run for the historical data extracted from Alpaca with the API and explained in chapter 2

Results_test for file Scenarios - Part 1 - 4_18_static_close.xlsx Sample of file generated for scenario 18 using static scenario with close price for each financial transaction required.

Results_test for file Scenarios - Part 1 - 4_18_static_open.xlsx Sample of file generated for scenario 18 using static scenario with open price for each financial transaction required.

Results_test for file Scenarios - Part 1 - 4_18_variable_close.xlsx Sample of file generated for scenario 18 using variable scenario with close price for each financial transaction required.

Results_test for file Scenarios - Part 1 - 4_18_variable_open.xlsx Sample of file generated for scenario 18 using variable scenario with open price for each financial transaction required.

Test results for file Part 2 - 4.xlsx Sample of file generated for part 2 of 4 of the 96 scenarios including only the final transaction for each scenario.

Test summary for file Scenarios - Part 2 - 4.xlsx Sample of file generated for part 2 of 4 of the 96 scenarios which bring the best scenario results for a faster analysis.

5 Conclusions

The results show enough data to support the hypothesis and help non-experienced traders obtain more net gains or lose less net loss by trading multiple times and using the open price.

It is worth running more scenarios to understand better if the percentage can be improved and which circumstances or limitations need to be considered to obtain higher positive results.

On the other hand, it will be important to modify the model to guarantee the number of tickers selected in the optimization process stays as close as possible to the number of tickers initially provided as a parameter even when buying full stocks only. As well as if there could be any mechanism to reduce the time to optimize the tickers.

Regardless of how unpredictable and emotional the trading world is, the historical data will never be able to predict the final results. Nonetheless, the test cases used for this project confirm the hypothesis and 10 months of testing using Dow Jones indexes and analyzing the results reinforce the importance of adjusting the model to the market so the successful rate increase.

Ultimately, we have detailed these concepts as part of our overall analysis to confirm our hypothesis from a Data Science perspective. This analysis is intended to remain a 'black box' for our non-experienced investors, who should only interact with a user-friendly mobile application to input their data and view the results, as described in section 6.1 'Simple mobile application' which would be available to the general public, including our expert investors.

6 Future Work

Contents

It is important to remember that the main objective is to keep this project as simple as possible for the non-experienced investors. This simplicity helps to demystify investing, reducing the intimidation factor that can come with more intricate financial concepts. As a result, non-experienced investors are more likely to stay engaged and consistent with their investments.

Additionally, cost-effectiveness and reduced risk profile can lead to more stable returns over time. Non-experienced investors can benefit from these steady gains as they are less likely to be affected by the volatility and higher expenses associated with more sophisticated investment approaches.

Moreover, seeing gains from simple investments can be highly motivating. As non-experienced investors begin to experience positive returns, their understanding of how investing works improves, and they become more comfortable with the process. This initial success can lead to increased financial literacy and a greater willingness to explore more advanced investment strategies in the future.

The combination of simplicity, cost-effectiveness, and observable gains creates a positive feedback loop that supports continued investment and financial growth which combined with trust and consistency can boost the investment culture in others as they foster a sense of security, satisfaction and attachment.

6.1 Simple mobile application

While we have described in detail all the platforms, data and processes used from a Data Science perspective, we are still focusing on non-experienced investors who may not be familiar with technology either so ideally we should create a simple, friendly and efficient mobile application that,

- captures the basic data: # stocks, budget, investment type and period,
- provides the necessary reports on static and/or variable investment results,
- stores historical data for future AI suggestions.

6.2 Tickers in the portfolio

We have reviewed some options we can implement to keep the number of tickers in the portfolio as close as possible to the initial parameter provided by the non-experienced investor. We need to prioritize between the weights of each ticker, the number of tickers in the portfolio and the number of stocks.

Once these other options are included in the code and there are some tests run to compare, this can also be included as part of the parameters the non-experienced investor may need to provide and only if the results are favourable from a net gain perspective.

6.3 Code optimization

As we indicated the variable scenario introduces more processing time in the code that we can revisit to reduce the processing time to guarantee the minimum time to execute a transaction in the market regardless of the scenario we are running.

We need to measure the current code time, define the goal and plan to optimize the code.

Bibliography

Tradingview: Free stock charts, stock quotes and trade ideas. https://www.tradingview.com/, a. Accessed on Date.

Tradingview data download and api information. https://www.tradingview.com/widget-docs/faq/data/, b. Accessed on Date.

Tradingview: Definition. https://en.wikipedia.org/wiki/TradingView, c. Accessed on Date.

Tradingview price information. https://www.tradingview.com/pricing/, d. Accessed on Date.

Alpaca. Alpaca api platform. https://docs.alpaca.markets/docs/alpaca-api-platform, a. Accessed on Date.

Alpaca. Alpaca markets. https://app.alpaca.markets/, b. Accessed on Date.

Alpaca. What is alpaca. https://alpaca.markets/support/what-is-alpaca, c. Accessed on Date.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, 2004.

BrokerChooser. Brokerage fee calculator, n.d. URL https://brokerchooser.com/brokerage-fee-calculator. [Online; accessed June 7, 2024].

M. Cagan. *Stock Market 101: From Bull and Bear Markets to Dividends, Shares, and Margins—Your Essential Guide to the Stock Market*. Adams Media, kindle edition edition, 2016.

Python Software Foundation. Python: The python programming language. https://www.python.org/. Accessed on Date.

Investopedia. Trading session, 2024. URL https://www.investopedia.com/terms/t/tradingsession.asp. [Online; accessed 30-May-2024].

Investopedia. Optimal portfolio size, n.d. URL https://www.investopedia.com/ask/answers/05/optimalportfoliosize.asp. [Online; accessed June 7, 2024].

Investopedia. Investopedia, Year the website was last updated. URL https://www.investopedia.com/. Accessed on Date.

P. K. Jha. What makes python a brilliant choice for data analysis. https://shorturl.at/09zgb. Accessed on Date.

H. Markowitz. Portfolio selection, 1952. URL https://www.jstor.org/stable/2975974. [Online; accessed September 8, 2024].

NYSE. Hours & calendars, 2024. URL https://www.nyse.com/markets/hours-calendars. [Online; accessed 30-May-2024].

Rosmarsan. Optimized pathway for non-experienced investors, 2024. URL https://github.com/rosmarsan/Optimized-pathway.git. [Online; accessed 30-May-2024].

W. F. Sharpe. A theory of market equilibrium under conditions of risk, 1964. URL https://www.jstor.org/stable/2977928. [Online; accessed September 8, 2024].

S&P Dow Jones Indices. Dow jones global titans 50 index factsheet, n.d. URL https://www.spglobal.com/spdji/en/idsenhancedfactsheet/file.pdf?calcFrequency=M&force_download=true&hostIdentifier=48190c8c-42c4-46af-8d1a-0cd5db894797&indexId=1301588. [Online; accessed 30-May-2024].

StockAnalysis. Tradingview review. https://stockanalysis.com/article/tradingview-review/. Accessed on Date.

SuperiorCodeLabs. Top 10 best programming languages to learn in 2024. https://shorturl.at/bmHvr. Accessed on Date.

TradingView. How to see the deepest historical data. https://shorturl.at/yahJg. Accessed on Date.

TradingView. Dji, 2024. URL https://www.tradingview.com/symbols/TVC-DJI/. [Online; accessed 30-May-2024].

Wikipedia. Dow jones industrial average, 2024a. URL https://en.wikipedia.org/wiki/Dow_Jones_Industrial_Average. [Online; accessed 30-May-2024].

Wikipedia. Dow jones global titans 50, 2024b. URL https://en.wikipedia.org/wiki/Dow_Jones_Global_Titans_50. [Online; accessed 30-May-2024].

Index