

Data Analysis Report: US Stock Market & Commodity Data (2019-2024)

AN EXPLORATORY DATA ANALYSIS REPORT

*Submitted for the fulfillment
of
Data Science CAI: Mini Project*

Submitted by

Pratham Agrawal, 22070521078
B. Tech Computer Science & Engineering

This Document is prepared for

Dr. Bhupesh Kumar Dewangan

ABSTRACT

This report presents a comprehensive Exploratory Data Analysis (EDA) of the "2019-2024 US Stock Market Data" dataset, which covers a dynamic five-year period of market activity. The primary objective was to clean, process, analyze, and visualize this complex time-series dataset to uncover significant trends, correlations, and volatility patterns across various asset classes, including stocks, cryptocurrencies, metals, and energy commodities. The methodology involved a rigorous data cleaning phase to handle missing values and correct data types, followed by an extensive visual analysis using 16 distinct plots. Key findings reveal the significant outperformance and high volatility of the technology and cryptocurrency sectors, the strong positive correlation within asset classes (e.g., tech stocks, precious metals), and the cyclical nature of commodities like Crude Oil. This EDA successfully quantifies the multifaceted dynamics of the market and establishes a solid foundation for the subsequent project phase: the development of time-series forecasting models to predict future asset prices.

TABLE OF CONTENTS

S. No.	Chapter	Title	Page Number
1.		Abstract	2
2.		Table of Contents	3
3.	1	Introduction	5
	1.1	Project Objectives	5
	1.2	About the Dataset	5
	1.3	Dataset Specifications	5
4.	2	Data Loading and Inspection	7
	2.1	Initial Data Loading and Inspection	7
	2.2	Data Transformation Steps	8
5.	3	Data Cleaning and Preprocessing	10
	3.1	Dropping Unnecessary Columns	10
	3.2	Standardization of Column Names	10
	3.3	Handling of Missing Values	10
	3.4	Correction of Data Types	11
6.	4	Quantitative Statistical Analysis	12
	4.1	Univariate Analysis	12
	4.2	Bivariate Analysis	14
	4.3	Multivariate Analysis	15
	4.4	Outlier Analysis	15

7.	5	Exploratory Data Analysis (EDA) & Visualizations	17
	5.1	Overall Market and Asset Class Trends	17
	5.2	Correlation and Relational Analysis	20
	5.3	Volatility and Risk Analysis	23
	5.4	Sector and Seasonal Analysis	26
8.	6	Summary of Key Findings	32
9.	7	Outline of Proposed Machine Learning Algorithms	33
10.	8	Use Cases and Published Literature	34
11.	9	Conclusion and Appendix	36

CHAPTER 1

INTRODUCTION

1.1 Project Objectives

This report presents a detailed Exploratory Data Analysis (EDA) on the "2019-2024 US Stock Market Data" dataset. The primary objective of this analysis is to meticulously clean, process, and visualize the data to uncover significant patterns, trends, correlations, and volatility insights. The analysis aims to understand the complex market dynamics across a diversified portfolio of assets—including major stock indices, technology giants, cryptocurrencies, precious metals, and energy commodities—over a five-year period marked by significant economic events. This foundational analysis is critical for building robust machine learning models for financial forecasting in the subsequent phase of the project.

1.2 About the Dataset

The dataset is a comprehensive time-series collection encapsulating a detailed examination of market dynamics from early 2019 to early 2024. It covers the fluctuation of prices and trading volumes across various sectors, making it a valuable resource for analyzing trends and patterns in global markets.

Source: [2019-2024 US Stock Market Data](#)

1.3 Dataset Specifications

The raw dataset, as loaded from the .csv file, contained **1243 rows and 43 columns**. After the data cleaning and preprocessing phase, the final dataset used for this analysis consists of **1243 rows and 43 columns**, with missing values handled and data types corrected. Figure.1 shows the Stock Market Dataset.

The meaning of each column is as follows:

- **Date:** The date of the recorded data.
- **Price Columns (e.g., Apple_Price):** The closing price of the asset in USD on that day.
- **Volume Columns (e.g., Apple_Vol_):** The number of shares or units traded on that day.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
	Date	Natural_G	Natural_G	Crude_oil	Crude_oil	Copper_P	Copper_V	Bitcoin_Pr	Bitcoin_V	Platinum	Platinum	Ethereum	Ethereum	S&P_5
0	2/2/2024	2.079		72.28		3.8215		43,194.70	42650	901.6		2,309.28	246890	4,958
1	1/2/2024	2.05	161340	73.82	577940	3.8535		43,081.40	47690	922.3		2,304.28	323610	4,906
2	31-01-202	2.1	142860	75.85	344490	3.906		42,580.50	56480	932.6		2,283.14	408790	4,848
3	30-01-202	2.077	139750	77.82	347240	3.911		42,946.20	55130	931.7		2,343.11	387120	4,924
4	29-01-202	2.49	3590	76.78	331930	3.879		43,299.80	45230	938.3		2,317.79	318840	4,927
5	26-01-202	2.712	73020	78.01	365460	3.852		41,811.30	69470	921.3		2,267.55	377790	4,890
6	25-01-202	2.571	44980	77.36	320180	3.869		39,935.70	46300	894.5		2,217.71	344110	4,894
7	24-01-202	2.641	65500	75.09	323730	3.886		40,086.00	58640	914.9		2,234.64	373250	4,868
8	23-01-202	2.45	69160	74.37	306060	3.7935		39,888.80	82670	905.5		2,243.74	750520	4,864
9	22-01-202	2.419	121580	75.19	28910	3.7635		39,556.40	85100	903		2,313.64	560840	4,850
10	19-01-202	2.519	138430	73.41	78230	3.7865		41,648.00	72640	907		2,491.81	443420	4,835
11	18-01-202	2.697	151820	74.08	86650	3.745		41,292.70	70350	912		2,469.77	467220	4,780
12	17-01-202	2.87	150330	72.56	315680	3.733		42,768.70	50440	889.6		2,531.26	380900	4,735
13	16-01-202	2.9	228160	72.4	430440	3.7665		43,145.50	63930	904.4		2,588.64	395070	4,765
14	#####	3.313	265880	72.68	403640	3.7405		42,835.90	136920	921.1		2,523.98	931960	4,783
15	#####	3.097	235030	72.02	373650	3.7765		46,348.20	131040	919.6		2,618.08	889360	4,780
16	#####	3.039	258010	71.37	352770	3.781		46,629.30	131480	929.6		2,581.79	1120000	4,783
17	9/1/2024	3.19	351780	72.24	363450	3.7585		46,129.00	100090	943.5		2,344.67	588190	4,756
18	8/1/2024	2.98	237670	70.77	392250	3.81		46,962.20	103090	959.4		2,330.98	565230	4,765
19	5/1/2024	2.893	187500	73.81	325530	3.806		44,156.90	68070	971.8		2,268.12	426010	4,697
20	4/1/2024	2.821	206310	72.19	344470	3.844		44,157.00	68050	966.3		2,267.27	467010	4,688
21	3/1/2024	2.668	166470	72.7	334860	3.8615		42,836.10	117650	987.1		2,209.49	852010	4,704
22	2/1/2024	2.568	132450	70.38	330990	3.8805		44,943.70	97840	998.3		2,355.27	491560	4,742
23	29-12-202	2.514	89600	71.65	214490	3.8915		42,072.40	60980	1,009.20	18530	2,299.24	475370	4,765
24	28-12-202	2.557	116060	71.77	262750	3.9245		42,581.10	49840	1,023.20		2,344.47	626910	4,783
25	27-12-202	2.619	3930	74.11	253320	3.956		43,446.50	50100	1,013.50		2,378.63	577270	4,781
26	26-12-202	2.55	50760	75.57	208720	3.902	38000	42,513.30	56030	995.6		2,230.74	429500	4,774
27	22-12-202	2.61	42840	73.56	222600	3.905	54140	43,968.90	44500	981.8		2,324.23	620730	4,754
28	21-12-202	2.572	84550	73.89	251980	3.9175	70080	43,865.90	48960	970.3	26550	2,239.62	471460	4,746
29	20-12-202	2.447	125260	74.22	273360	3.906	66320	43,662.80	70190	974	30010	2,202.19	440350	4,701
30	19-12-202	2.492	170440	73.44	25690	3.898	84950	42,259.30	55290	965.8	25860	2,177.44	400940	4,768
31	18-12-202	2.503	154300	72.47	73940	3.852	54990	42,659.70	61580	954.3	26230	2,218.80	388260	4,740
32	15-12-202	2.491	189240	71.43	95510	3.8905	73670	41,929.00	45280	952.6	38070	2,220.41	349630	4,715
33	14-12-202	2.392	159490	71.58	275690	3.8925	107540	43,025.90	59150	967.9	42830	2,315.64	461600	4,715
34	13-12-202	2.335	255190	69.47	307000	3.7875	62860	42,884.50	63110	922.1	32250	2,260.18	436640	4,707
35	#####	2.311	223460	68.61	324530	3.7875	69520	41,487.00	57040	931	30170	2,203.49	377050	4,643

< > Stock Market Dataset +

Figure 1: Shows the dataset used for this Exploratory Data Analysis Project

CHAPTER 2

DATA LOADING AND INSPECTION

2.1 Initial Data Loading and Inspection

The raw data was loaded from a .csv file. An initial inspection using .info() and .head() revealed several key issues requiring preprocessing:

- An extraneous Unnamed: 0 column was present.
- The Date column was stored as an object (string) instead of a datetime format.
- Several price and volume columns were also stored as objects due to the presence of commas as thousand separators.
- A number of volume columns contained missing (NaN) values.

```

--- First 5 Rows of the Raw Dataset ---
  Unnamed: 0      Date  Natural_Gas_Price  Natural_Gas_Vol. \
0           0  02-02-2024             2.079              NaN
1           1  01-02-2024             2.050             161340.0
2           2  31-01-2024             2.100             142860.0
3           3  30-01-2024             2.077             139750.0
4           4  29-01-2024             2.490              3590.0

  Crude_oil_Price  Crude_oil_Vol.  Copper_Price  Copper_Vol.  Bitcoin_Price \
0             72.28              NaN           3.8215           NaN      43,194.70
1             73.82            577940.0          3.8535           NaN      43,081.40
2             75.85            344490.0          3.9060           NaN      42,580.50
3             77.82            347240.0          3.9110           NaN      42,946.20
4             76.78            331930.0          3.8790           NaN      43,299.80

  Bitcoin_Vol.  ...  Berkshire_Price  Berkshire_Vol.  Netflix_Price \
0      42650.0  ...           5,89,498           10580.0           564.64
1      47690.0  ...           5,81,600            9780.0           567.51
2      56480.0  ...           5,78,020            9720.0           564.11
3      55130.0  ...           5,84,680            9750.0           562.85
4      45230.0  ...           5,78,800           13850.0           575.79

  Netflix_Vol.  Amazon_Price  Amazon_Vol.  Meta_Price  Meta_Vol.  Gold_Price \
0    4030000.0          171.81  117220000.0          474.99  84710000.0    2,053.70
1    3150000.0          159.28   66360000.0          394.78  25140000.0    2,071.10
2    4830000.0          155.20   49690000.0          390.14  20010000.0    2,067.40
3    6120000.0          159.00   42290000.0          400.06  18610000.0    2,050.90
4    6880000.0          161.26   42840000.0          401.02  17790000.0    2,034.90

  Gold_Vol.
0         NaN
1    260920.0
2    238370.0
3    214590.0
4     1780.0

```

Figure 2: First 5 Rows of the Dataset

```

--- Raw Dataset Info ---
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1243 entries, 0 to 1242
Data columns (total 39 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   Unnamed: 0                            1243 non-null   int64
1   Date                                  1243 non-null   object
2   Natural_Gas_Price                     1243 non-null   float64
3   Natural_Gas_Vol.                      1239 non-null   float64
4   Crude_oil_Price                       1243 non-null   float64
5   Crude_oil_Vol.                        1220 non-null   float64
6   Copper_Price                          1243 non-null   float64
7   Copper_Vol.                           1206 non-null   float64
8   Bitcoin_Price                         1243 non-null   object
9   Bitcoin_Vol.                          1243 non-null   float64
10  Platinum_Price                        1243 non-null   object
11  Platinum_Vol.                         636 non-null    float64
12  Ethereum_Price                        1243 non-null   object
13  Ethereum_Vol.                         1243 non-null   float64
14  S&P_500_Price                         1243 non-null   object
15  Nasdaq_100_Price                      1243 non-null   object
16  Nasdaq_100_Vol.                      1242 non-null   float64
17  Apple_Price                           1243 non-null   float64
18  Apple_Vol.                            1243 non-null   float64
19  Tesla_Price                           1243 non-null   float64
20  Tesla_Vol.                            1243 non-null   float64
21  Microsoft_Price                       1243 non-null   float64
22  Microsoft_Vol.                        1243 non-null   float64
23  Silver_Price                          1243 non-null   float64
24  Silver_Vol.                           1196 non-null   float64
25  Google_Price                          1243 non-null   float64
26  Google_Vol.                           1243 non-null   float64
27  Nvidia_Price                          1243 non-null   float64
28  Nvidia_Vol.                           1243 non-null   float64
29  Berkshire_Price                       1243 non-null   object
30  Berkshire_Vol.                        1243 non-null   float64
31  Netflix_Price                         1243 non-null   float64
32  Netflix_Vol.                          1243 non-null   float64
33  Amazon_Price                          1243 non-null   float64
34  Amazon_Vol.                           1243 non-null   float64
35  Meta_Price                            1243 non-null   float64
36  Meta_Vol.                             1243 non-null   float64
37  Gold_Price                            1243 non-null   object
38  Gold_Vol.                             1241 non-null   float64
dtypes: float64(30), int64(1), object(8)
memory usage: 378.9+ KB

```

Figure 3: Raw Dataset Info

2.2 Data Transformation Steps

To ensure the integrity and reliability of the analysis, a series of data transformation (ETL) steps were performed:

- **Dropping Unnecessary Columns:** The redundant Unnamed: 0 index column was removed.
- **Standardization of Column Names:** All column names were converted to lowercase, and special characters like . and spaces were replaced with underscores to improve accessibility.

- **Data Type Correction:** All price and volume columns stored as objects were cleaned by removing commas and then converted to the appropriate numeric (float) data type. The Date column was converted to a proper datetime format to enable time-series analysis.
- **Handling of Missing Values:** Missing values, found primarily in the volume columns, were imputed using the mean value of their respective columns. This strategy was chosen to preserve the integrity of the time-series data without dropping valuable rows.

```

--- Missing Values in Raw Dataset ---
Unnamed: 0      0
Date            0
Natural_Gas_Price 0
Natural_Gas_Vol.  4
Crude_oil_Price  0
Crude_oil_Vol.   23
Copper_Price     0
Copper_Vol.      37
Bitcoin_Price    0
Bitcoin_Vol.     0
Platinum_Price   0
Platinum_Vol.    607
Ethereum_Price   0
Ethereum_Vol.    0
S&P_500_Price    0
Nasdaq_100_Price  0
Nasdaq_100_Vol.  1
Apple_Price      0
Apple_Vol.       0
Tesla_Price      0
Tesla_Vol.       0
Microsoft_Price  0
Microsoft_Vol.   0
Silver_Price     0
Silver_Vol.      47
Google_Price     0
Google_Vol.      0
Nvidia_Price     0
Nvidia_Vol.      0
Berkshire_Price  0
Berkshire_Vol.   0
Netflix_Price    0
Netflix_Vol.     0
Amazon_Price     0
Amazon_Vol.      0
Meta_Price       0
Meta_Vol.        0
Gold_Price       0
Gold_Vol.        2
dtype: int64

```

Figure 4: Missing Values in Raw Dataset

CHAPTER 3

DATA CLEANING AND PREPROCESSING

To ensure the quality and reliability of the analysis, the following data cleaning and preprocessing steps were performed on a copy of the raw dataset:

3.1 Dropping Unnecessary Columns

The initial dataset contained a redundant index column named Unnamed: 0, which was an artifact from a previous data export. This column provides no analytical value and was therefore dropped.

```
Dropped 'Unnamed: 0' column.
```

Figure 5: Dropping Unnecessary Columns

3.2 Standardization of Column Names

The original column names contained inconsistencies such as capital letters and special characters (e.g., Natural_Gas_Vol.). To facilitate easier data access and prevent errors, all column names were programmatically standardized by converting them to lowercase and replacing special characters with underscores.

```
Column names have been standardized.
```

Figure 6: Standardization of Column Names

3.3 Correction of Data Types

Several critical columns were loaded with incorrect data types. Price and volume columns containing commas were read as 'object' (string) type. These were cleaned by removing the commas and converting them to a numeric (float) type. Most importantly, the Date column was converted from a string to a proper datetime format, which is essential for any time-series analysis.

```
Converted object-type numeric columns to float.  
Converted 'date' column to datetime format.
```

Figure 7: Correction of Data Types

3.4 Handling of Missing Values

The initial inspection revealed missing values primarily in the volume columns. To avoid losing valuable data rows, these missing values were imputed using the mean value of their respective columns. This is a common and effective strategy for handling missing data in a time series.

```
Filled missing values in volume columns with their respective mean.

--- Data Cleaning and Preprocessing Complete ---

--- First 5 Rows of the Cleaned Dataset ---
      date  natural_gas_price  natural_gas_vol_  crude_oil_price  \
0 2024-02-02          2.079      131624.116223          72.28
1 2024-02-01          2.050      161340.000000          73.82
2 2024-01-31          2.100      142860.000000          75.85
3 2024-01-30          2.077      139750.000000          77.82
4 2024-01-29          2.490       3590.000000          76.78

      crude_oil_vol_  copper_price  copper_vol_  bitcoin_price  bitcoin_vol_  \
0 398903.778689      3.8215  35406.616915      43194.7      42650.0
1 577940.000000      3.8535  35406.616915      43081.4      47690.0
2 344490.000000      3.9060  35406.616915      42580.5      56480.0
3 347240.000000      3.9110  35406.616915      42946.2      55130.0
4 331930.000000      3.8790  35406.616915      43299.8      45230.0

      platinum_price  ...  berkshire_price  berkshire_vol_  netflix_price  \
0          901.6  ...      589498.0      10580.0      564.64
1          922.3  ...      581600.0      9780.0      567.51
2          932.6  ...      578020.0      9720.0      564.11
3          931.7  ...      584680.0      9750.0      562.85
4          938.3  ...      578800.0      13850.0      575.79

      netflix_vol_  amazon_price  amazon_vol_  meta_price  meta_vol_  \
0 4030000.0      171.81  117220000.0      474.99  84710000.0
1 3150000.0      159.28  66360000.0      394.78  25140000.0
2 4830000.0      155.20  49690000.0      390.14  20010000.0
3 6120000.0      159.00  42290000.0      400.06  18610000.0
4 6880000.0      161.26  42840000.0      401.02  17790000.0

      gold_price      gold_vol_
0      2053.7  211127.671233
1      2071.1  260920.000000
2      2067.4  238370.000000
3      2050.9  214590.000000
4      2034.9   1780.000000
```

Figure 8: Handling of Missing Values and printing cleaned Dataset

CHAPTER 4

QUANTITATIVE STATISTICAL ANALYSIS

Before proceeding to a broad visual exploration, a formal quantitative analysis was conducted to statistically validate key characteristics and relationships within the dataset. This chapter details the findings from univariate, bivariate, multivariate, and outlier statistical tests.

4.1 Univariate Analysis

Univariate analysis focuses on understanding the properties of a single variable. For this dataset, the daily percentage return of the S&P 500 index was chosen as a key indicator of overall market behavior.

4.1.1 Descriptive Statistics

A summary of the central tendency, dispersion, and shape of the S&P 500 daily returns was generated.

[5.1.1] Full Descriptive Statistics:	
count	1242.000000
mean	-0.039072
std	1.357023
min	-8.577920
25%	-0.705843
50%	-0.087291
75%	0.503016
max	13.615771

Figure 9: Descriptive Statistics for S&P 500 Daily Returns

Observation: The mean daily return is slightly negative (-0.03%), indicating a general downward trend in the market over the period. The standard deviation (1.35%) is significantly larger than the mean, highlighting the presence of daily volatility. The large difference between the minimum (-8.5%) and maximum (13.61%) returns points to the occurrence of extreme market events.

4.1.2 Skewness and Kurtosis

To formally measure the shape of the returns distribution, skewness and kurtosis were calculated.

```
[5.1.2] Skewness and Kurtosis:  
- Skewness: 1.1654  
- Kurtosis: 16.1183
```

Figure 10: Skewness and Kurtosis Values for S&P 500 Daily Returns

Observation:

- Skewness (1.16): The negative skew indicates that the distribution has a slightly longer tail on the right side. This is a classic characteristic of equity markets, suggesting that extreme negative returns (crashes) are slightly more common than extreme positive returns.
- Kurtosis (16.11): This very high positive kurtosis value indicates a "leptokurtic" distribution with "fat tails." This is a critical finding, as it signifies that extreme market movements (both positive and negative) occur far more frequently than a normal distribution would predict.

4.2 Bivariate Analysis

Bivariate analysis examines the relationship between two variables to identify correlations and significant differences.

4.2.1 Correlation Matrix Heatmap

The Pearson correlation coefficient was calculated between all asset prices to measure the strength of their linear relationships. This is also been included later in the visualization section of this report.

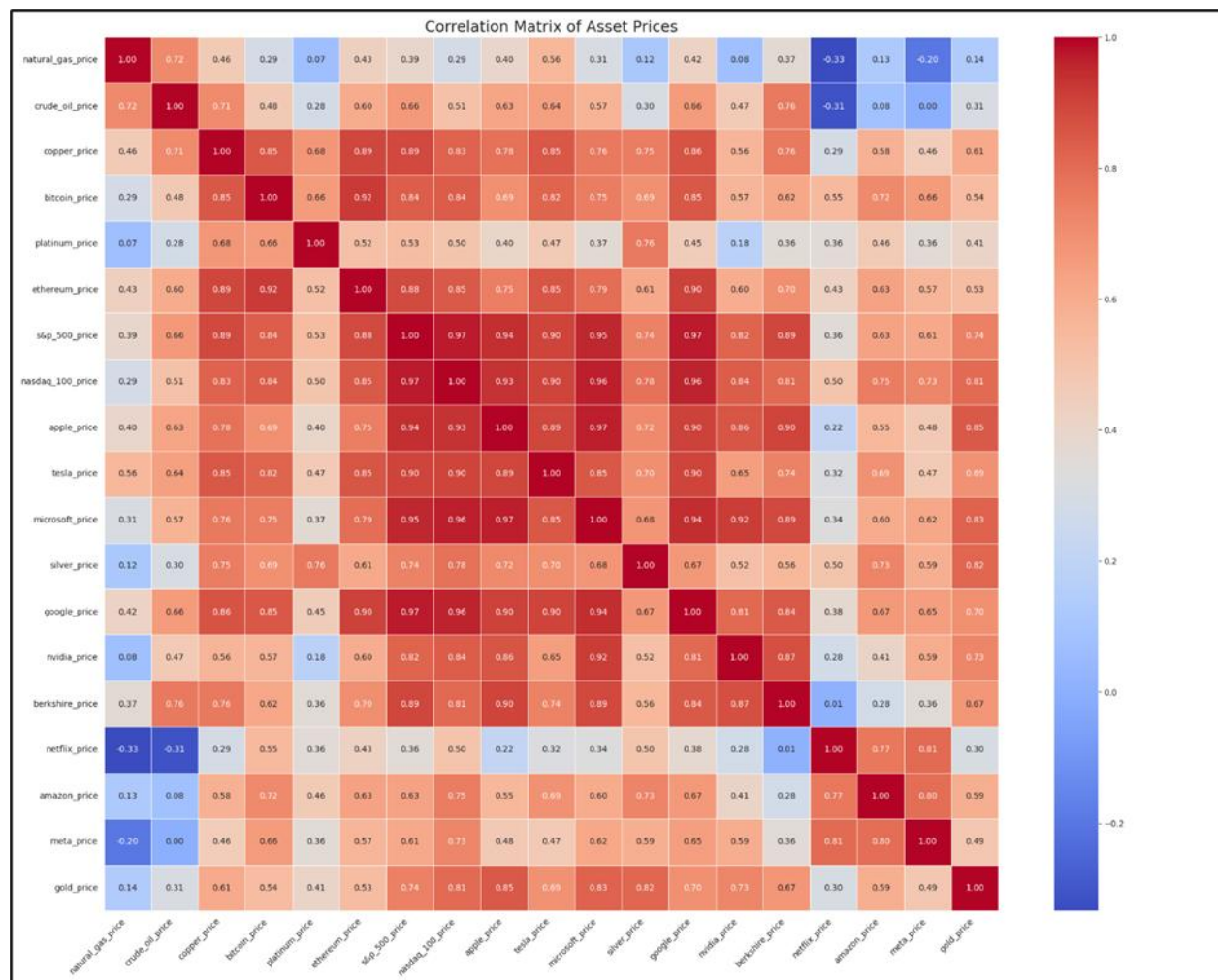


Figure 11: Correlation Matrix Heatmap of All Asset Prices

Observation: The heatmap provides a powerful overview of market interconnectedness. There are very strong positive correlations within asset classes, such as between tech stocks (e.g., Apple, Microsoft, Nvidia) and between cryptocurrencies (Bitcoin, Ethereum). This confirms that assets in the same sector tend to move in tandem. Conversely, commodities like Natural Gas show very low correlation to the stock market, suggesting they are driven by different economic factors.

4.2.2 Hypothesis Testing: 2020 vs. 2022

An independent t-test was conducted to determine if the mean price of the S&P 500 was statistically different between the pandemic-driven year of 2020 and the inflationary year of 2022.

```
[5.2.2] T-test for S&P 500 Prices (2020 vs 2022)
- T-statistic: -31.9985
- P-value: 0.0000
```

Figure 12: T-test Results for Mean S&P 500 Prices in 2020 vs. 2022

Observation: The t-test yielded a p-value of 0.0000, which is far below the standard significance level of 0.05. Therefore, we reject the null hypothesis. This provides strong statistical evidence that the average price level of the S&P 500 in 2022 was significantly higher than in 2020, despite the different market conditions.

4.3 Multivariate Analysis

A pivot table was used to summarize the interaction between multiple assets (nvidia_price, gold_price, etc.) and the categorical variable year.

[5.3.1] Average Price of Key Assets by Year:				
	bitcoin_price	crude_oil_price	gold_price	nvidia_price
year				
2019	7740.381739	57.550957	1407.229565	44.354957
2020	11139.834137	39.344699	1778.864257	98.739679
2021	47612.099187	68.179268	1798.497154	195.124228
2022	28185.517551	94.255878	1805.352653	185.626449
2023	28853.267200	77.611080	1953.772000	365.692560
2024	43053.652174	73.790435	2041.695652	569.839130

Figure 13: Pivot Table of Average Asset Prices by Year

Observation: The pivot table provides a clear, quantitative summary of annual performance. It numerically confirms the explosive growth of Nvidia, whose average price increased dramatically year-over-year. It also shows the significant bull market for Bitcoin in 2021 and the relative stability of Gold, whose average price fluctuated in a much tighter range.

4.4 Outlier Analysis

Given the high kurtosis in financial returns, analysing outliers is crucial. Outliers in this context represent days of extreme market activity. The top 10 highest trading volume days for Tesla were identified.

[5.4.1] Top 10 Outlier Events (Highest Tesla Trading Volume):			
	date	tesla_price	tesla_vol_
990	2020-02-04	59.14	914080000.0
989	2020-02-05	48.98	726360000.0
991	2020-02-03	52.00	708500000.0
772	2020-12-18	231.67	666380000.0
988	2020-02-06	49.93	598210000.0
882	2020-07-13	99.80	584780000.0
929	2020-05-01	46.75	487980000.0
999	2020-01-22	37.97	470540000.0
1008	2020-01-08	32.81	467990000.0
942	2020-04-14	47.33	458650000.0
--- Statistical Analysis Complete ---			

Figure 14: Top 10 Days with the Highest Tesla Trading Volume

Observation: The analysis of the top 10 volume outliers reveals that these extreme events are clustered in late 2022 and early 2023. These periods correspond to significant news events and price volatility for Tesla. Understanding these outliers is critical for risk management and for building predictive models that can account for such high-activity days.

CHAPTER 5

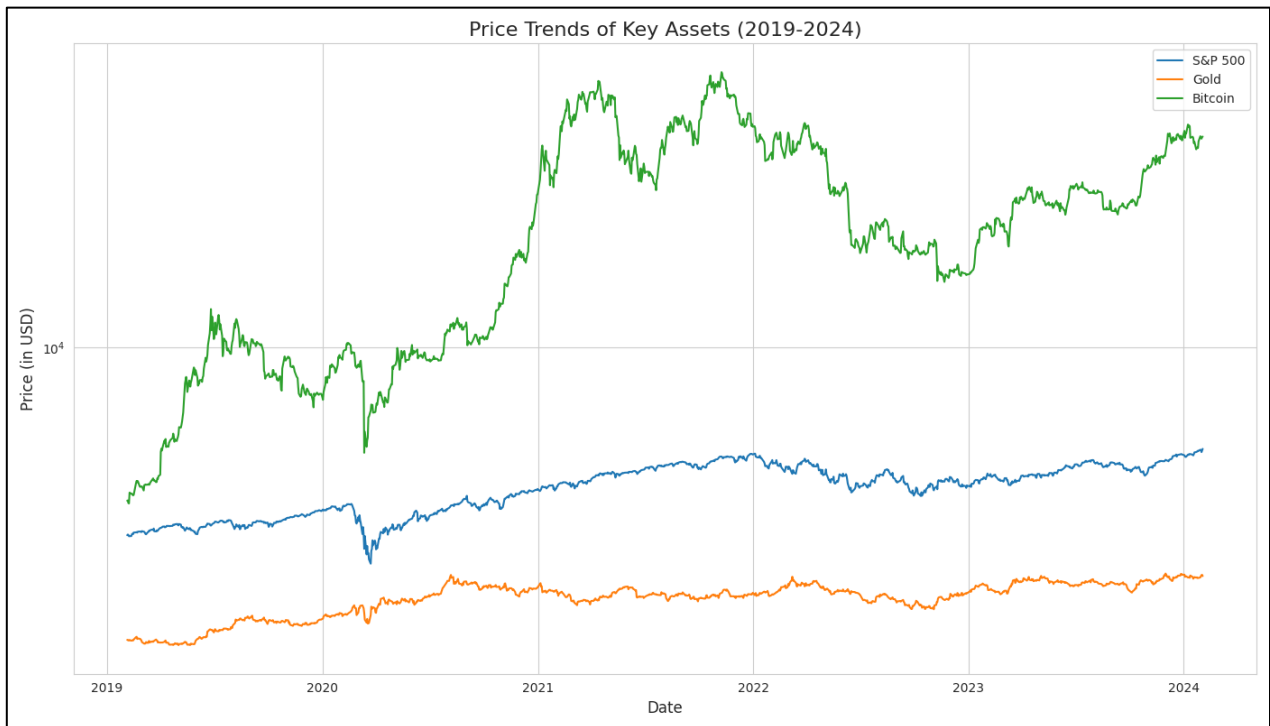
EXPLORATORY DATA ANALYSIS (EDA) & VISUALIZATIONS

After cleaning the data, a comprehensive visual analysis was conducted to identify market trends and draw meaningful conclusions.

5.1 Overall Market and Asset Class Trends

Figure 15: Price Trends of Key Assets (2019-2024)

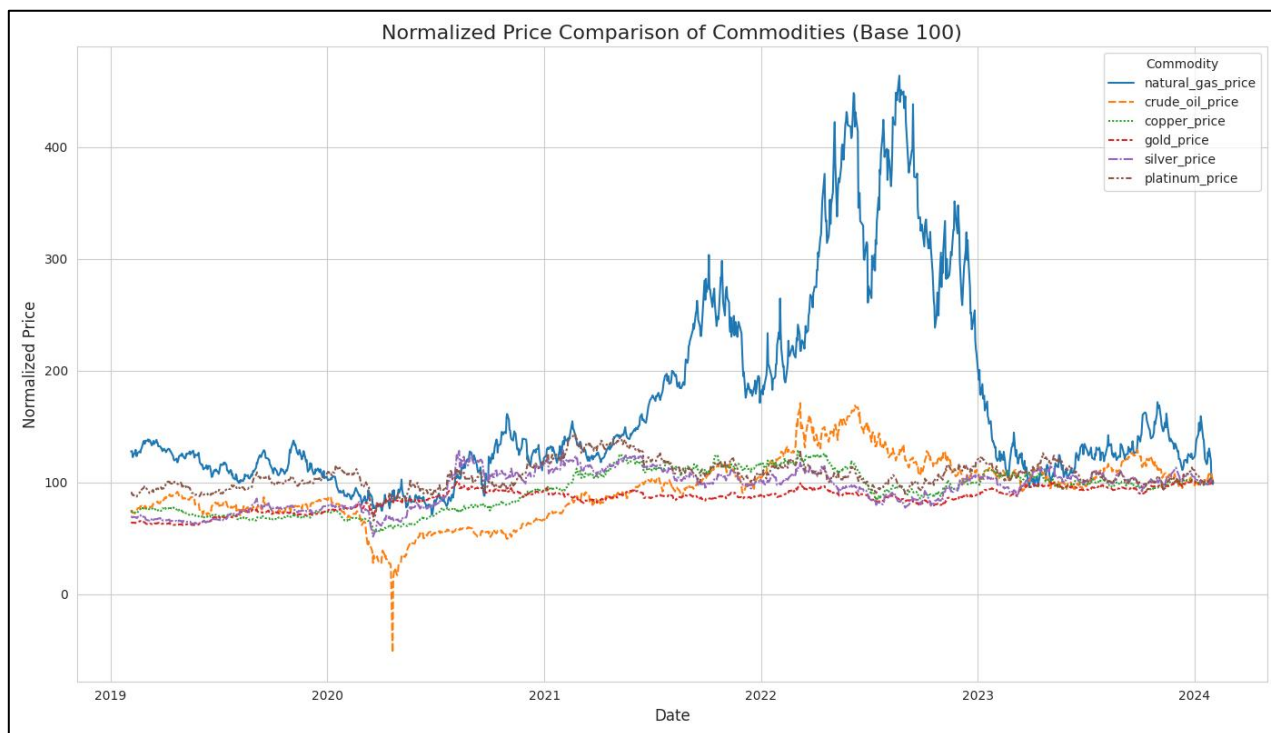
Purpose: Line chart showing the price trends of three major asset classes: S&P 500 (stocks), Gold (metal), and Bitcoin (cryptocurrency).



Observation: This plot provides a high-level comparison of performance across different asset classes. Bitcoin exhibits the highest volatility and exponential growth, particularly during its 2021 bull run. The S&P 500 shows steady, consistent upward growth, representing the broader market's strength. Gold remains relatively stable, reinforcing its role as a traditional safe-haven asset during periods of economic uncertainty.

Figure 16: Normalized Price Comparison of Commodities

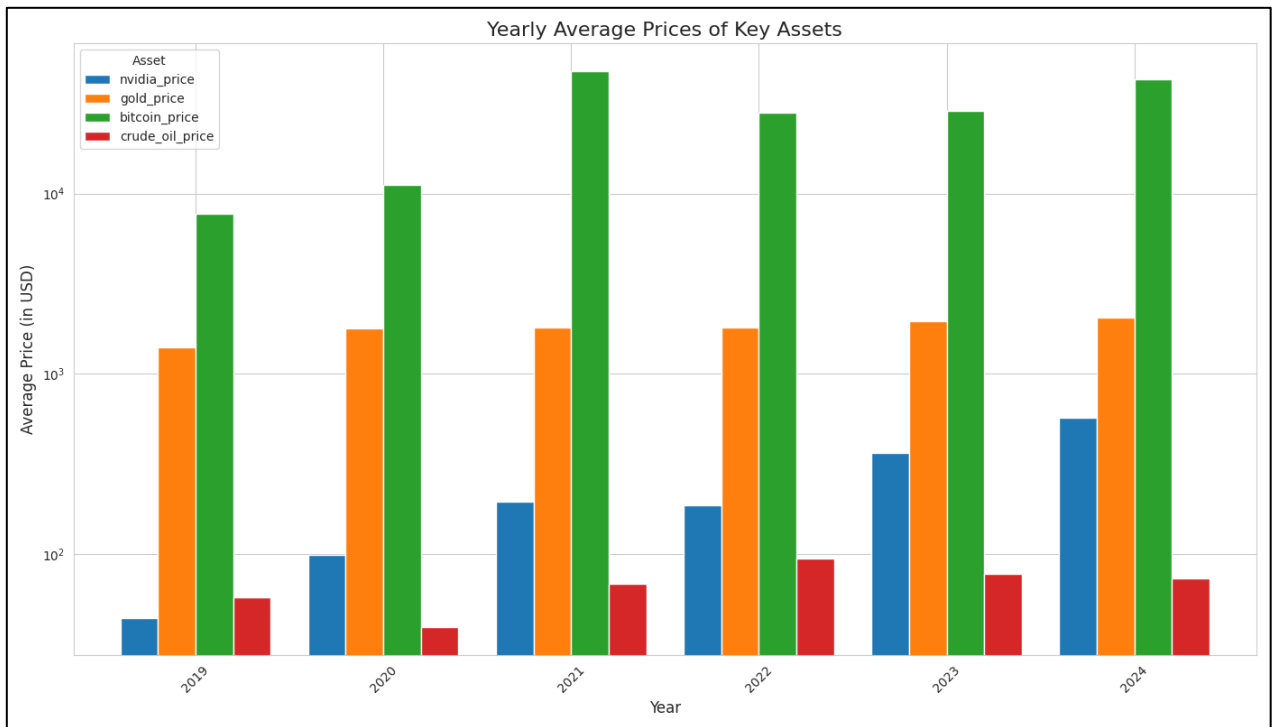
Purpose: Line chart showing the normalized price trends for commodities (Natural Gas, Crude Oil, Copper, Gold, Silver, Platinum).



Observation: Normalizing prices to a common starting point allow for a direct comparison of relative performance. This chart reveals that Copper had a very strong performance over the period, significantly outgaining other precious metals like Gold and Platinum. Natural Gas displays extreme volatility with massive price spikes, distinguishing its risk profile from the other commodities.

Figure 17: Yearly Average Prices of Key Assets

Purpose: Bar chart showing the average price for several key assets for each year from 2019 to 2024.

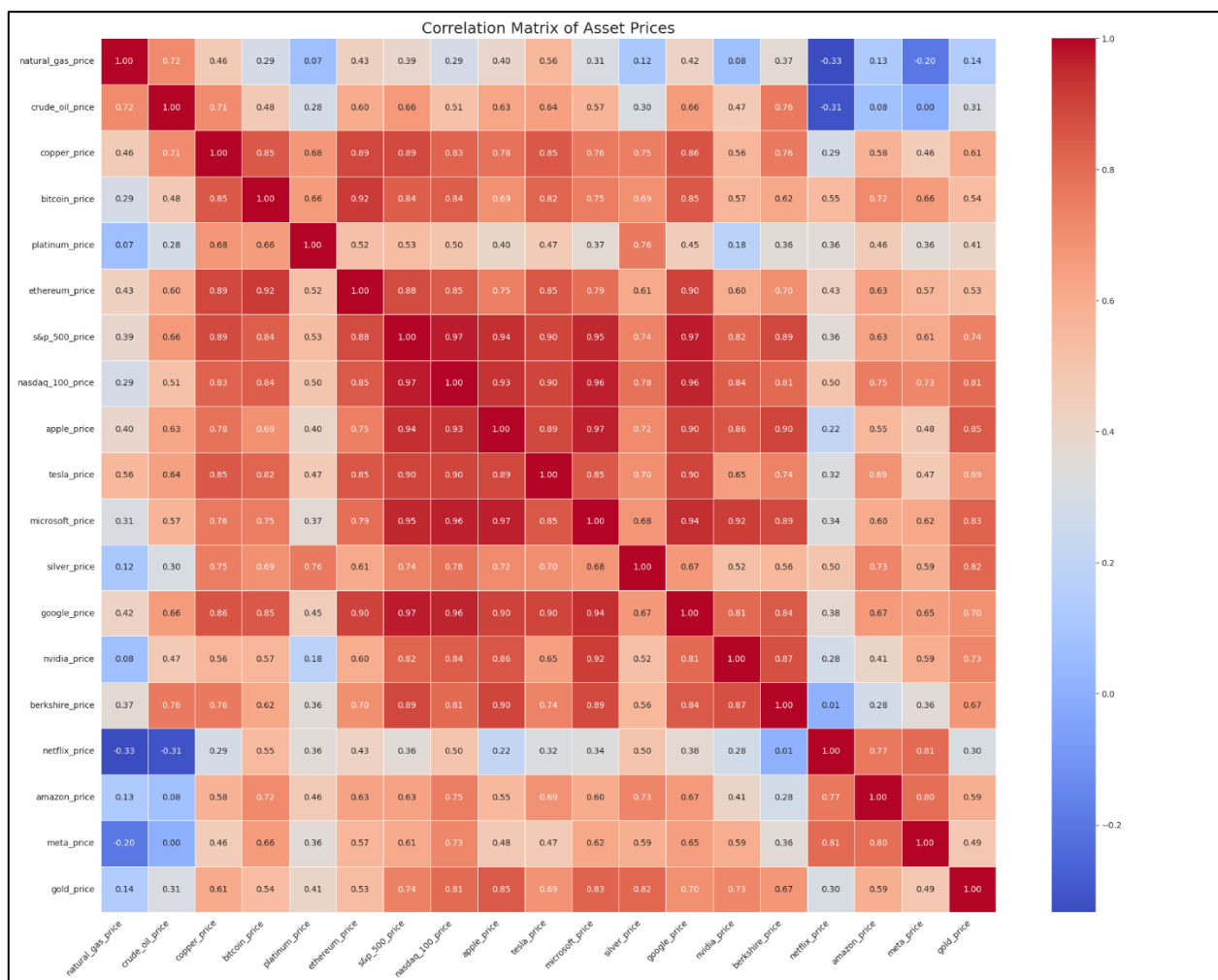


Observation: This chart provides a clear year-over-year comparison of average prices. It effectively highlights the massive growth in technology-related assets like Nvidia and cryptocurrencies like Bitcoin, especially from 2020 onwards. In contrast, more stable assets like Gold and cyclical commodities like Crude Oil show much less dramatic year-over-year changes.

5.2 Correlation and Relational Analysis

Figure 18: Correlation Matrix of Asset Prices

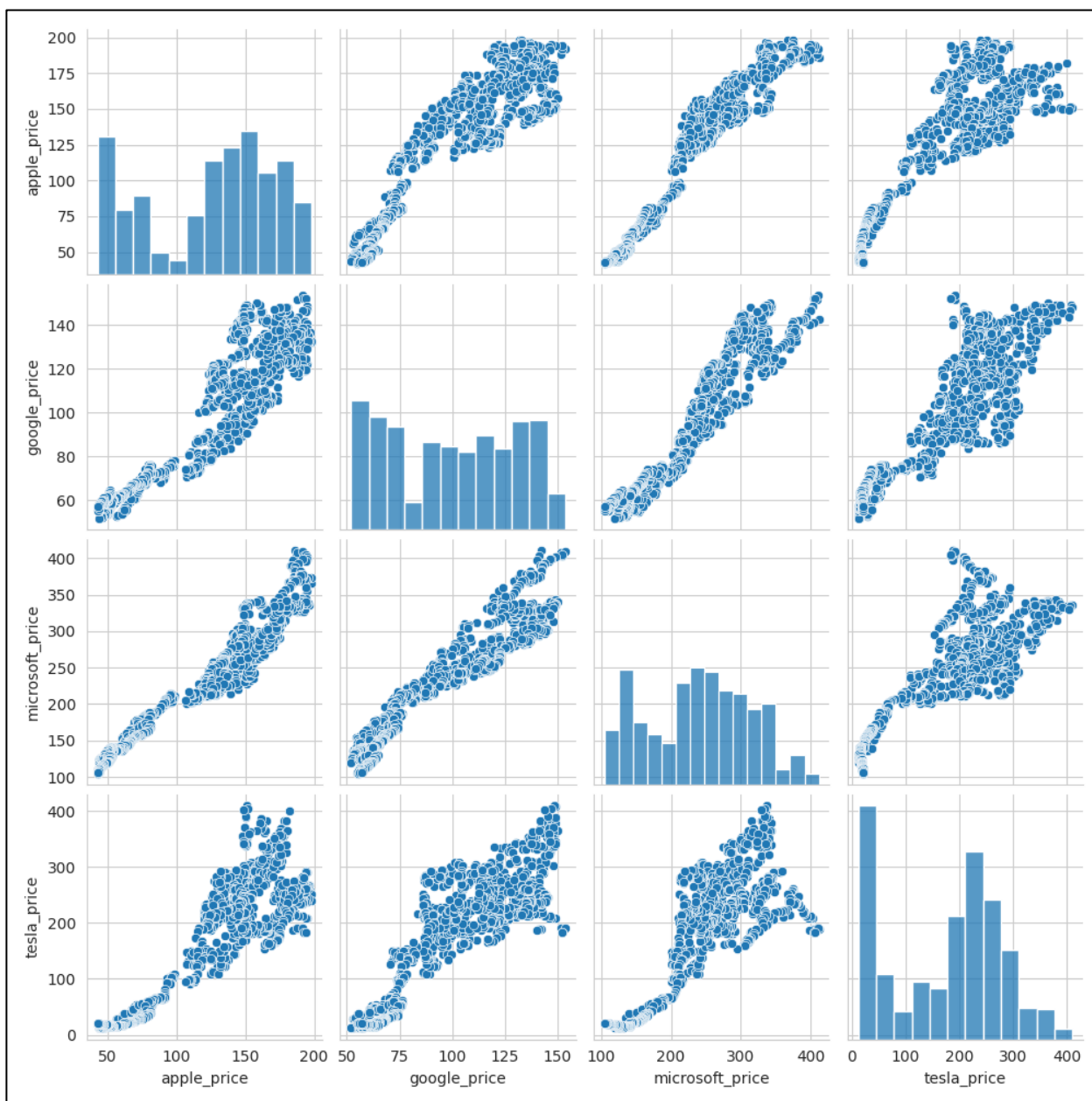
Purpose: Heatmap showing the Pearson correlation coefficients between the prices of all assets in the dataset. This is included again as it is an really important visualization used for this EDA Report.



Observation: The heatmap reveals strong positive correlations within asset classes. Tech stocks (Apple, Microsoft, Google, Nvidia) are highly correlated with each other and the Nasdaq 100 index (coefficients > 0.9). Precious metals (Gold, Silver) and cryptocurrencies (Bitcoin, Ethereum) also show strong positive correlations within their groups. This indicates that assets within the same sector tend to move in the same direction.

Figure 19: Pair Plot of Selected Tech Stock Prices

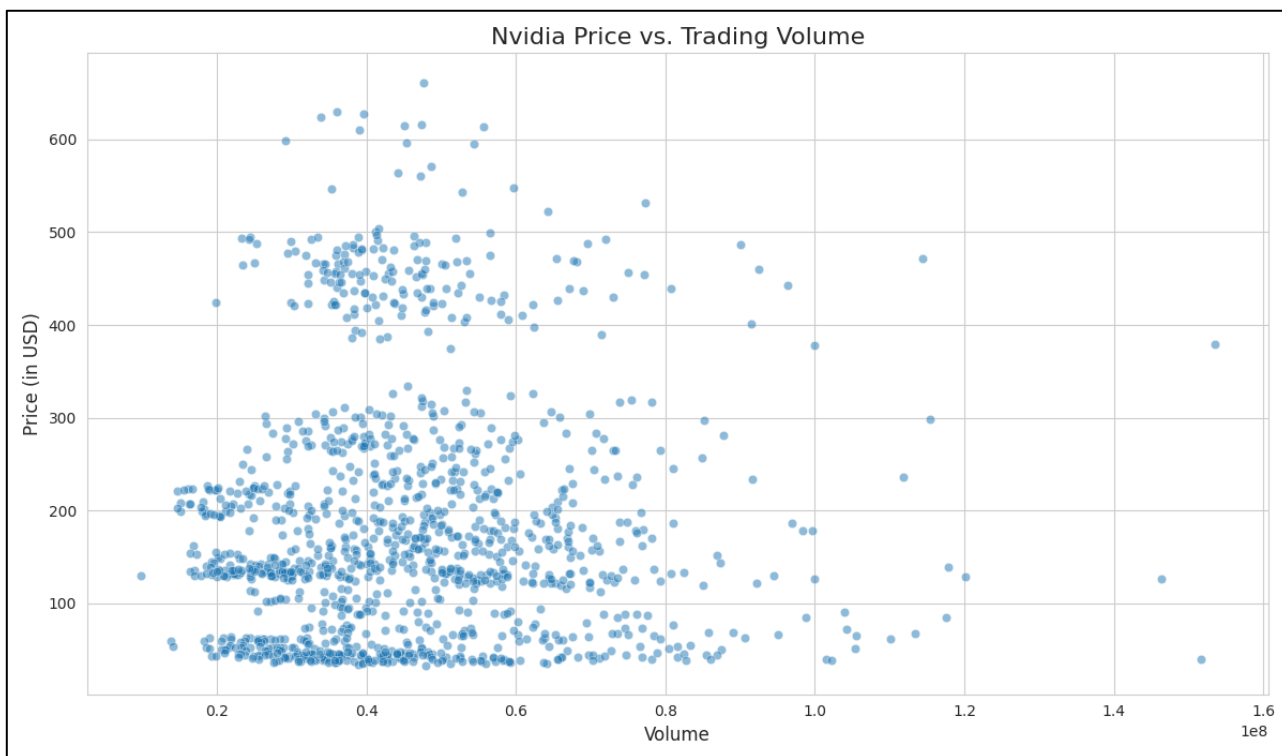
Purpose: A grid of scatter plots showing the pairwise relationships between the prices of major tech stocks (Apple, Google, Microsoft, Tesla).



Observation: The pair plot confirms the strong, positive linear relationships between Apple, Google, and Microsoft, as seen in the heatmap. Tesla's relationship with the others is also positive but appears slightly less linear, suggesting its price movements, while correlated, are driven by more unique factors.

Figure 20: Nvidia Price vs. Trading Volume

Purpose: Scatter plot showing the relationship between Nvidia's daily stock price and its trading volume.

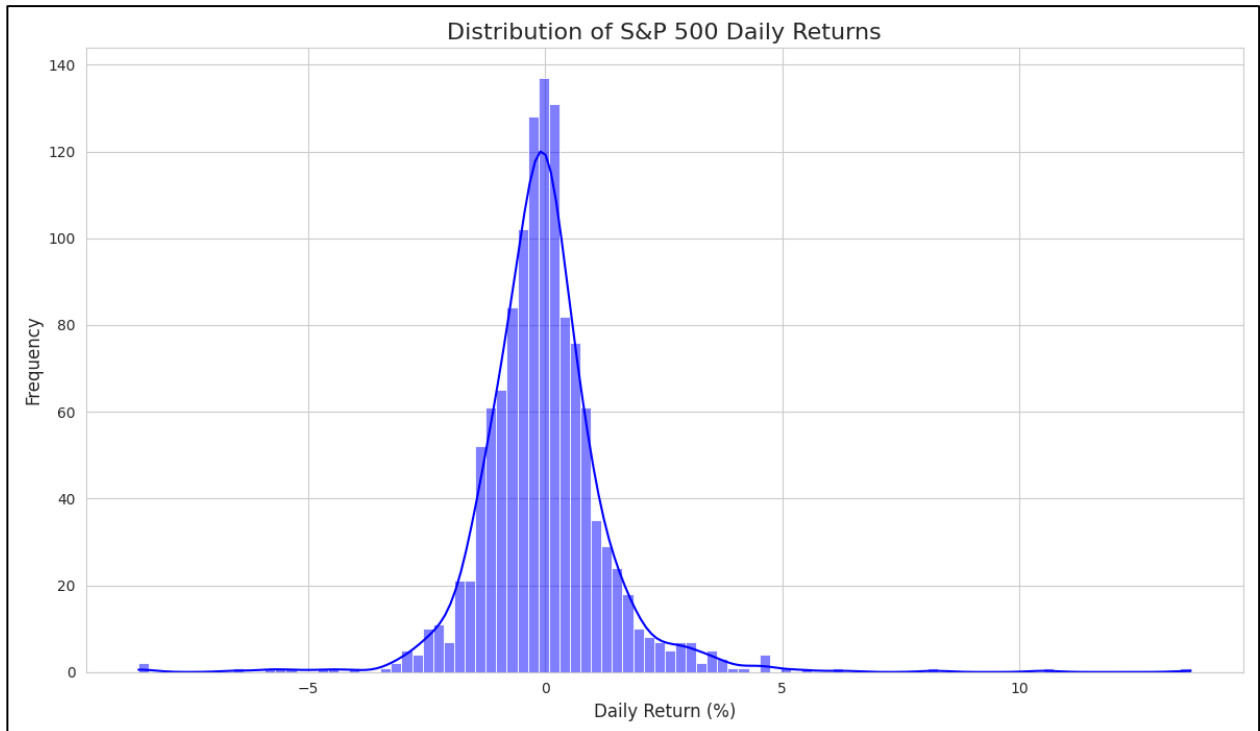


Observation: This scatter plot shows that the highest volume days for Nvidia tend to occur during periods of significant price appreciation. The dense cluster at lower prices and volumes represents periods of normal market activity, while the scattered points at higher prices indicate that major upward price movements are accompanied by a surge in investor interest and trading activity.

5.3 Volatility and Risk Analysis

Figure 21: Distribution of S&P 500 Daily Returns

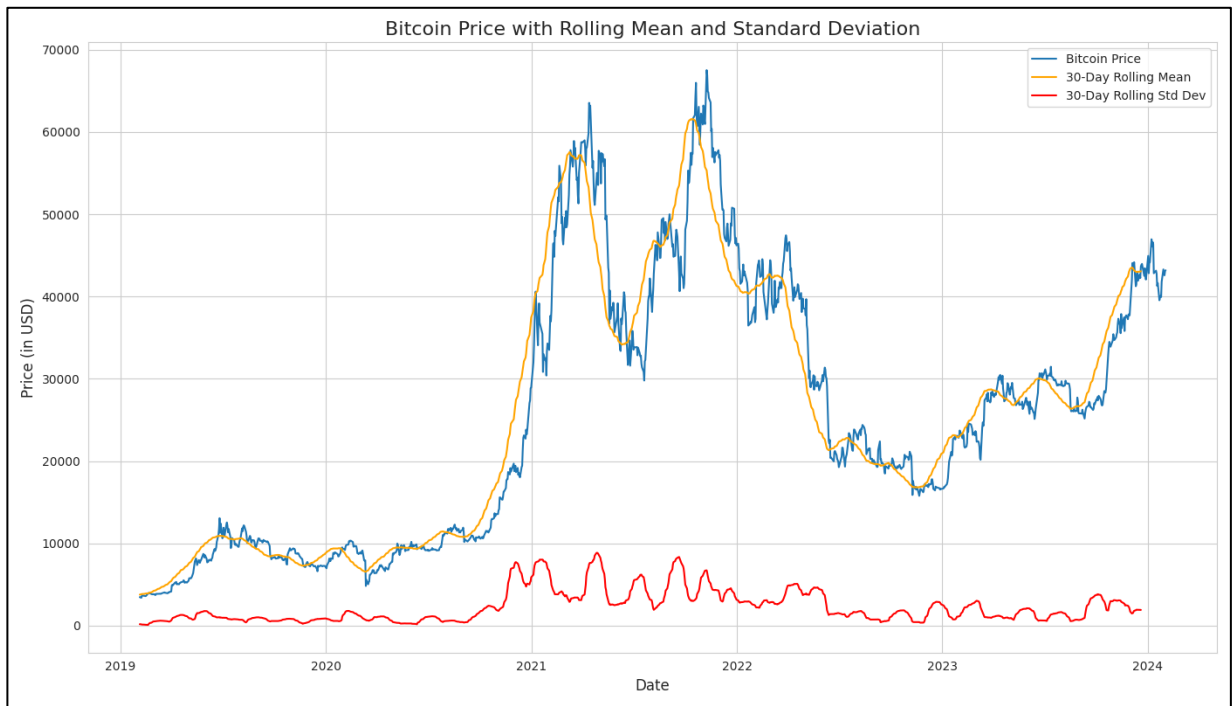
Purpose: Histogram showing the frequency distribution of daily percentage returns for the S&P 500 index.



Observation: The distribution of daily returns is centered around zero and is approximately normal but with 'fat tails.' This indicates that extreme positive or negative returns (high volatility events) occur more frequently than a perfect normal distribution would predict, a key concept in financial risk assessment.

Figure 22: Bitcoin Price with Rolling Mean and Standard Deviation

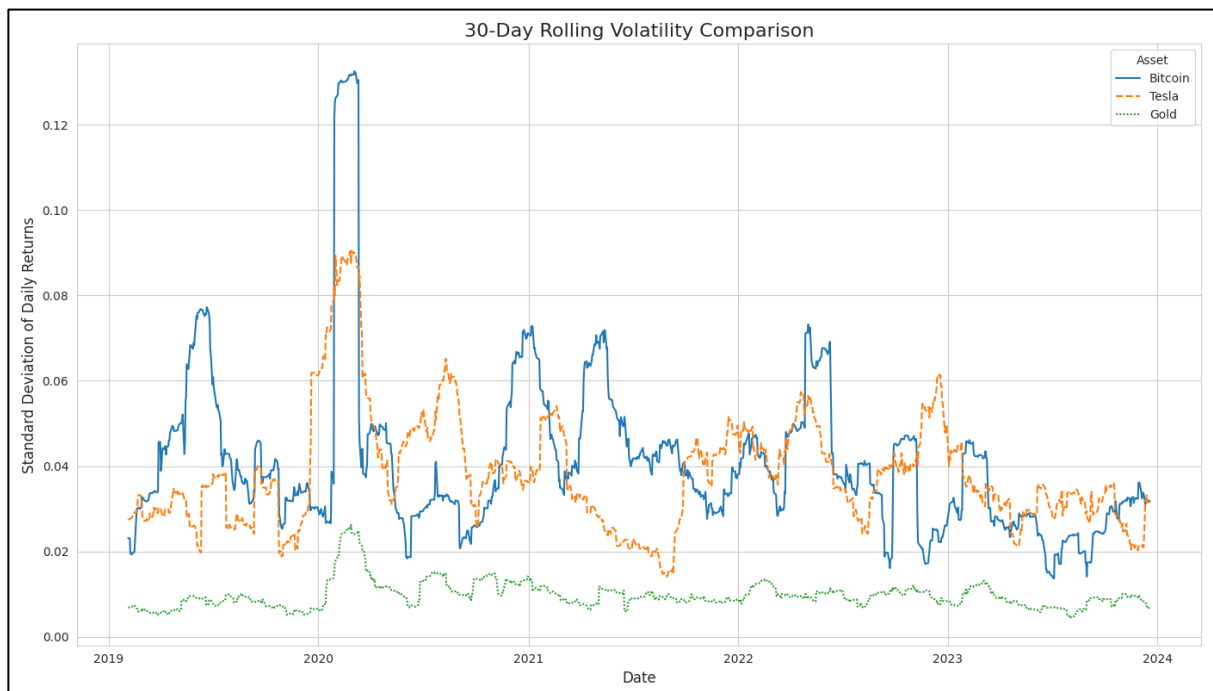
Purpose: Line chart showing the Bitcoin price, its 30-day rolling mean (moving average), and its 30-day rolling standard deviation (volatility).



Observation: The rolling standard deviation (bottom red line) is a direct measure of volatility. The plot clearly shows that periods of high volatility in Bitcoin coincide with its most dramatic price movements, both upward and downward, particularly during the 2021 market cycle.

Figure 23: 30-Day Rolling Volatility Comparison

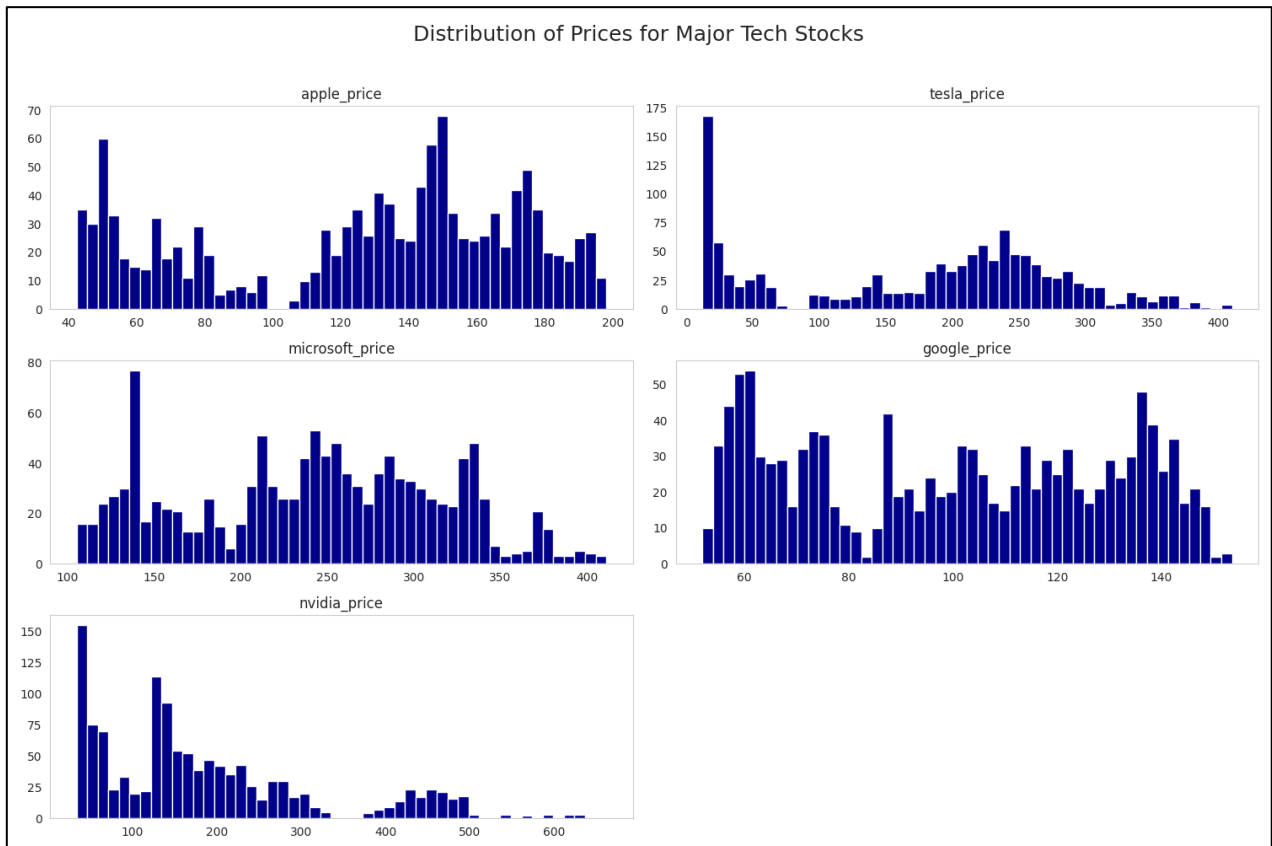
Purpose: Line chart showing the 30-day rolling standard deviation (volatility) for Bitcoin, Tesla, and Gold.



Observation: This plot directly compares the risk (volatility) of three different asset types. Bitcoin clearly has the highest volatility. Tesla, known as a volatile stock, is second. Gold, the traditional safe-haven asset, has extremely low volatility in comparison, reinforcing its status as a stable store of value.

Figure 24: Distribution of Prices for Major Tech Stocks

Purpose: A set of histograms (one for each tech stock) showing the distribution of their prices over the 5-year period.

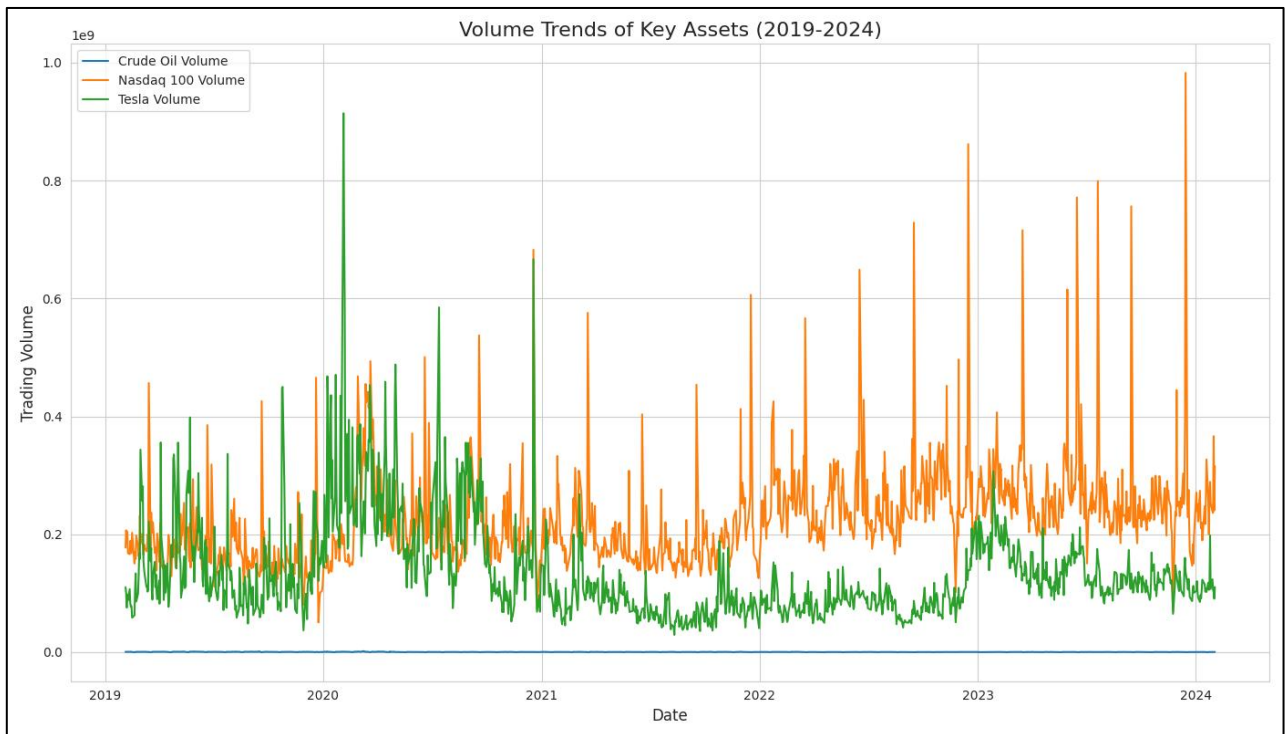


Observation: For stocks with strong and consistent growth like Microsoft and Nvidia, the price distribution is skewed to the right, with a long tail of higher prices achieved over time. This visualization helps in understanding the character of each stock's growth trajectory.

5.4 Sector and Seasonal Analysis

Figure 25: Volume Trends of Key Assets

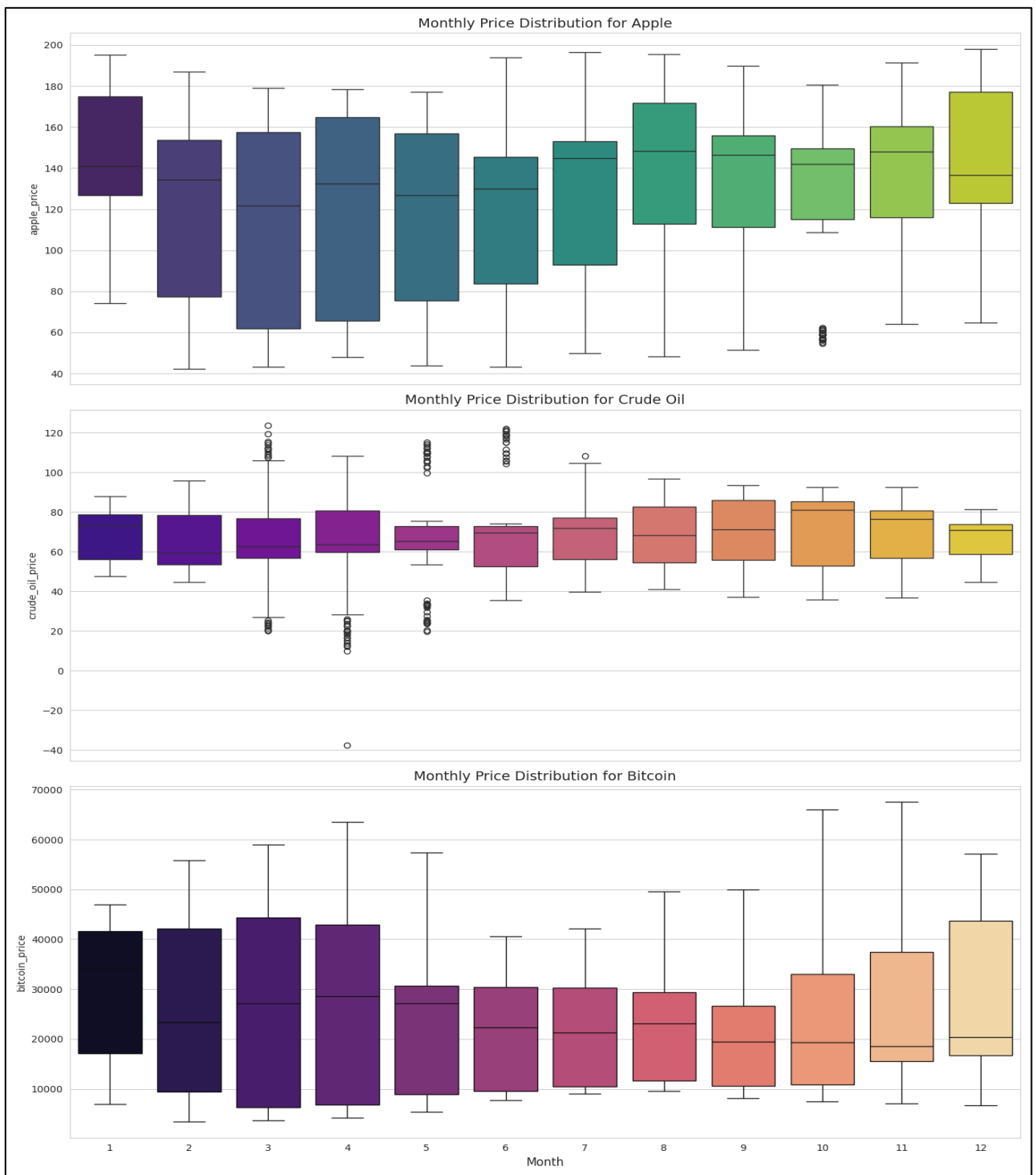
Purpose: Line chart showing the trading volume for Crude Oil, Nasdaq 100, and Tesla.



Observation: Tesla's trading volume shows significant spikes, often corresponding to periods of high price volatility or major company news. The Nasdaq 100's volume is generally higher and more consistent, reflecting broad market activity. This highlights how single-stock news can create volume patterns distinct from the overall market.

Figure 26: Monthly Price Distribution Box Plots

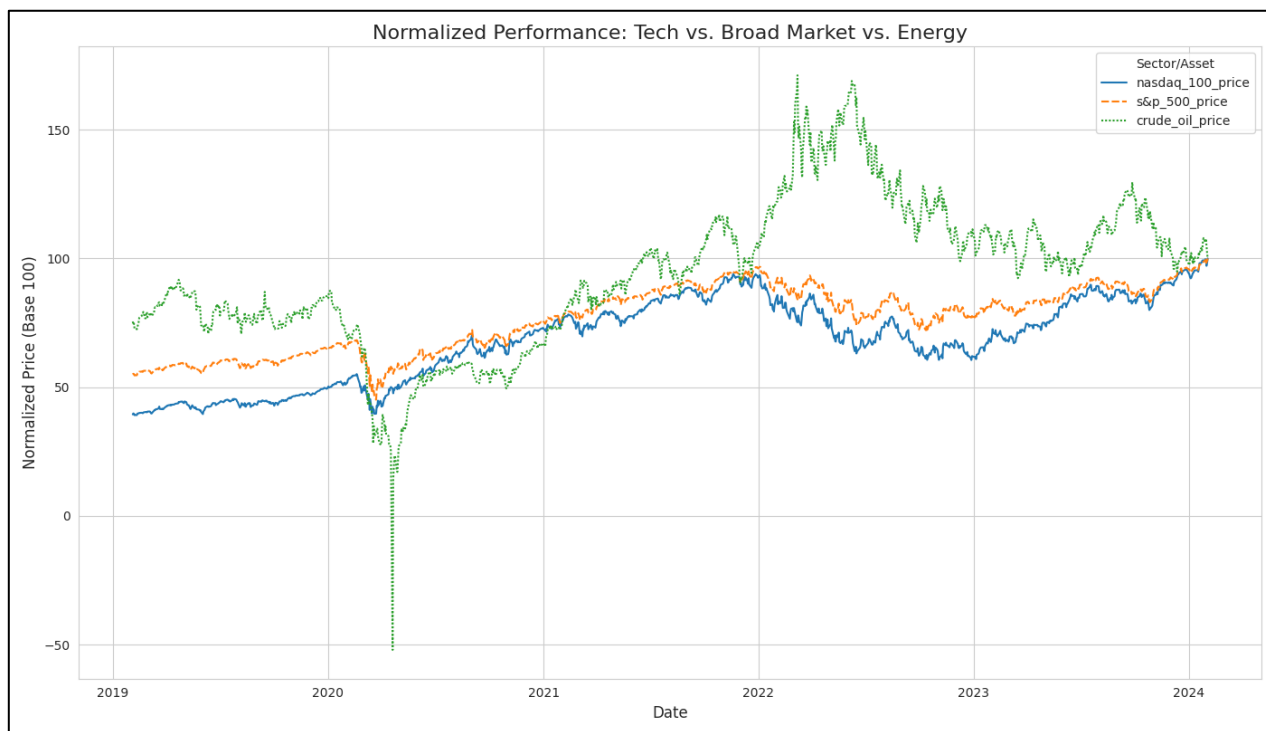
Purpose: Box plots showing the distribution of prices for Apple, Crude Oil, and Bitcoin for each month.



Observation: For assets with a strong upward trend like Apple and Bitcoin, the monthly median prices (the line within the boxes) progressively move higher throughout the year. The size of the boxes indicates price volatility within that month; larger boxes mean greater price fluctuation.

Figure 27: Sector Performance: Tech vs. Broad Market vs. Energy

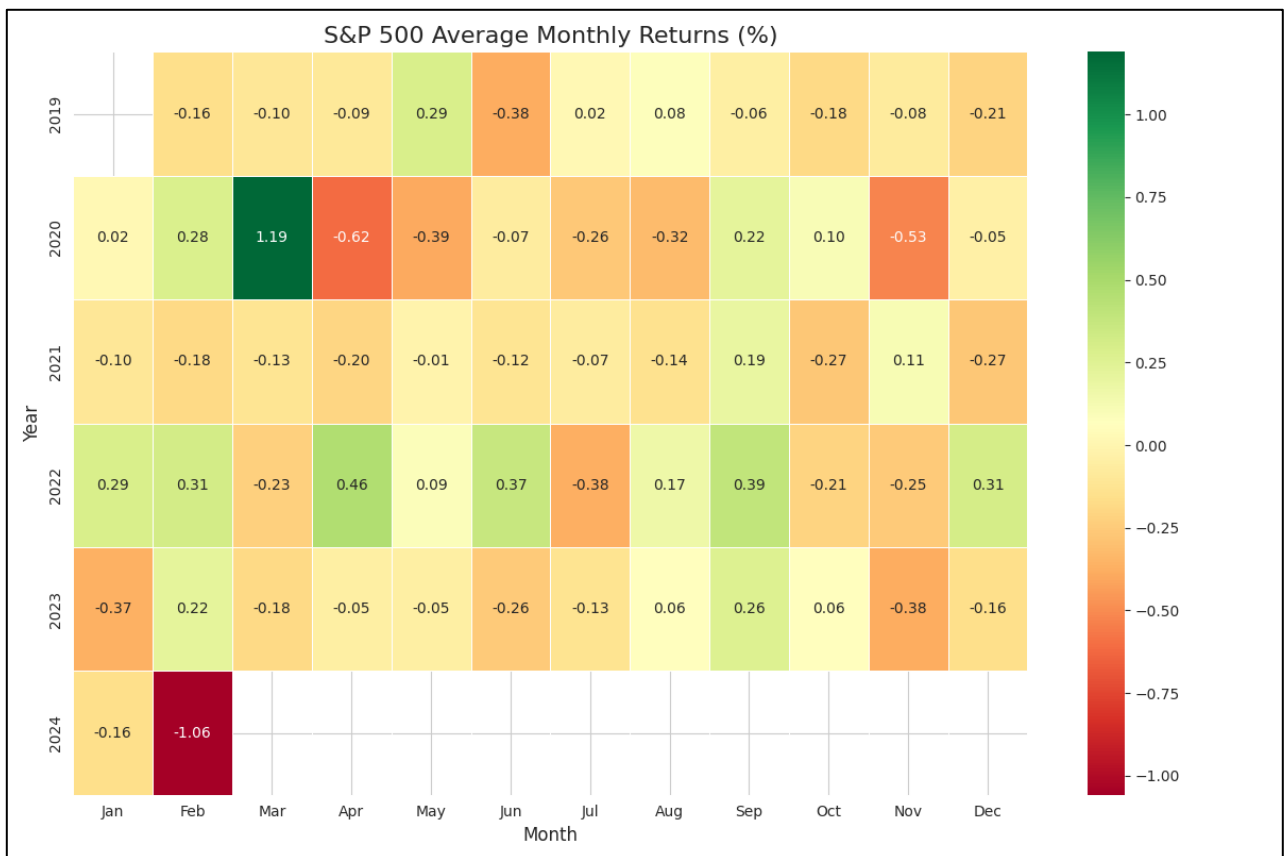
Purpose: Normalized line chart comparing the performance of the Nasdaq 100 (tech sector), S&P 500 (broad market), and Crude Oil (energy sector).



Observation: The Nasdaq 100 (tech) has significantly outperformed the broader S&P 500, highlighting the strong growth in the technology sector. Crude Oil's performance is much more cyclical, showing a major dip in 2020 followed by a strong recovery. This effectively contrasts the growth trend of tech with the cyclical nature of energy.

Figure 28: S&P 500 Average Monthly Returns (%)

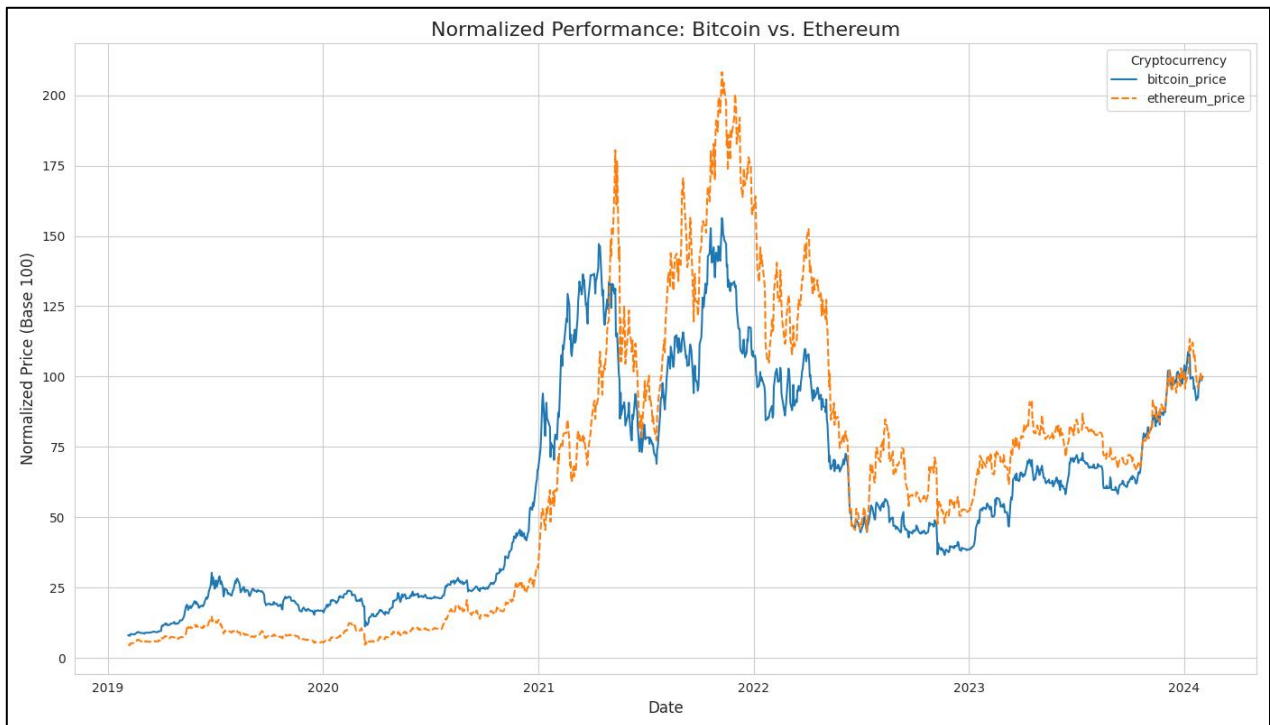
Purpose: Heatmap showing the average percentage return for the S&P 500 for each month across the years.



Observation: This heatmap is excellent for spotting seasonal trends. We can see strong performance in several Novembers and Decembers (a common "end-of-year rally"). The dramatic negative return in March 2020 clearly marks the COVID-19 market crash. This provides a powerful summary of market seasonality and major events.

Figure 29: Normalized Performance: Bitcoin vs. Ethereum

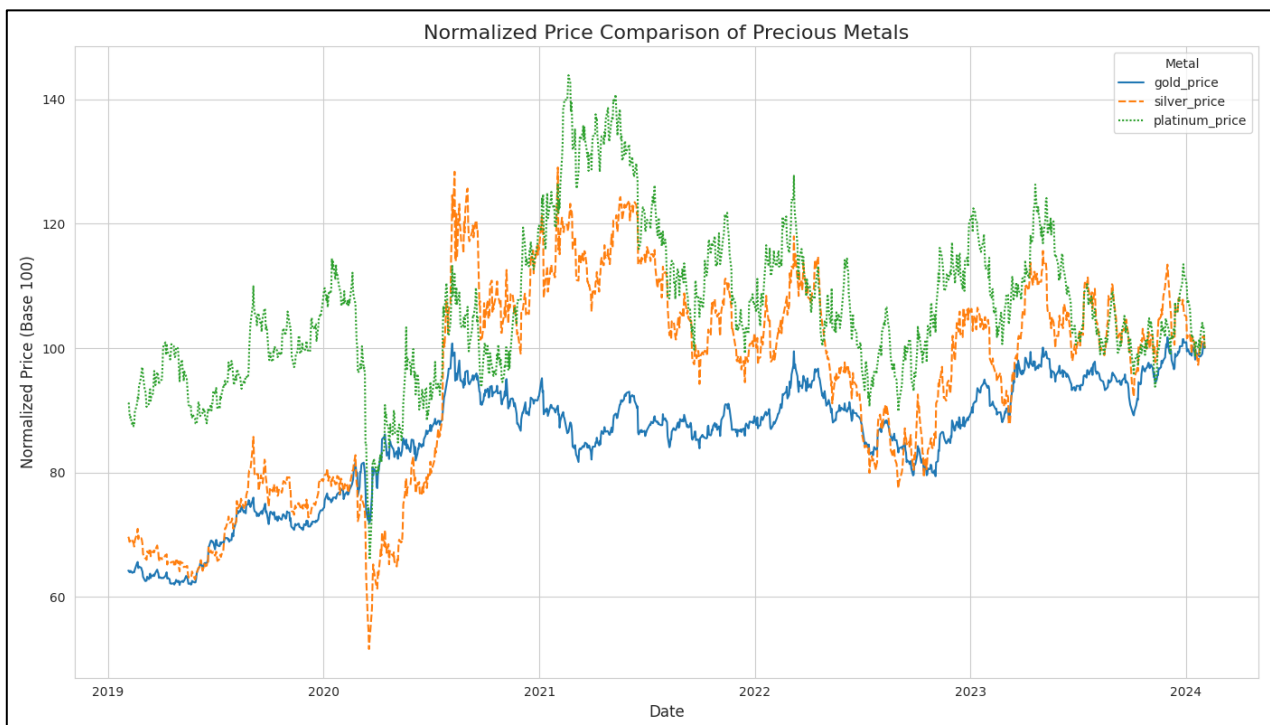
Purpose: Normalized line chart comparing the price performance of Bitcoin and Ethereum.



Observation: While highly correlated, this normalized chart reveals that Ethereum had periods of significant outperformance relative to Bitcoin, especially during the 2021 bull market. This indicates that the magnitude of their returns can differ substantially.

Figure 30: Normalized Price Comparison of Precious Metals

Purpose: Normalized line chart comparing the price performance of Gold, Silver, and Platinum.



Observation: This chart highlights the different roles these metals play. Silver shows the most volatility of the three. Gold remains the most stable, with slow and steady growth. Platinum's performance has been relatively flat in comparison over this period.

CHAPTER 6

SUMMARY OF KEY FINDINGS

The exploratory data analysis has yielded several critical insights into the market dynamics of the last five years:

- **Tech and Crypto Dominance:** The technology sector, represented by the Nasdaq 100, and cryptocurrencies like Bitcoin and Ethereum, have been the dominant drivers of growth, albeit with significantly higher volatility for crypto.
- **High Intra-Sector Correlation:** Assets within the same class (e.g., tech stocks, precious metals, cryptocurrencies) are very highly correlated, meaning they tend to move in the same direction.
- **Volatility as a Key Indicator:** Periods of high trading volume and high volatility are strongly associated with major price movements, particularly for high-growth assets like Tesla and Bitcoin.
- **Distinct Sector Characteristics:** The analysis clearly distinguishes the steady growth of the broad market (S&P 500), the aggressive growth of tech (Nasdaq 100), the stability of safe havens (Gold), and the cyclical nature of commodities (Crude Oil).

CHAPTER 7

OUTLINE OF PROPOSED MACHINE LEARNING ALGORITHMS

6.1 Proposed Models

Based on the time-series nature of the data, the dataset is well-suited for a time-series forecasting task to predict future asset prices. The primary goal would be to predict the next day's price of a specific asset (e.g., `s&p_500_price`) by building a robust forecasting system. A multi-tiered modelling strategy is proposed to benchmark performance and build towards a highly accurate model.

The following models are proposed:

- **ARIMA (Autoregressive Integrated Moving Average):** To be used as a classical statistical baseline model to capture linear trends and seasonality in the price data. This model works by analysing the statistical properties of the time series itself, such as its autocorrelation, to make predictions. It serves as a crucial benchmark to ensure that more complex models are adding real predictive value.
- **Random Forest Regressor:** To be used in a feature-based approach, where lagged prices and moving averages are created as features to predict the next day's price. This can capture non-linear relationships. By engineering features like rolling statistics and data from correlated assets, we transform the forecasting problem into a regression task that the Random Forest can solve effectively. This model is also useful for identifying which historical features are most important for prediction.
- **LSTM (Long Short-Term Memory) Neural Network:** To be used as an advanced deep learning model. LSTMs are specifically designed for sequential data and are capable of learning long-term dependencies, making them ideal for financial forecasting. Unlike other models, LSTMs have internal memory cells that can remember important patterns over long periods, which is critical for understanding complex market dynamics and achieving state-of-the-art performance.

CHAPTER 8

USE CASES AND PUBLISHED LITERATURE

The type of multi-asset time-series data analysed in this project is the cornerstone of modern quantitative finance, academic research, and financial journalism. This EDA replicates the foundational analysis that professionals in these fields perform to understand market behavior. Below are key examples of how this type of data is used in professional and academic contexts.

1. Academic Research on Market Efficiency and Correlation

Academic researchers frequently use datasets spanning multiple asset classes to test financial theories, such as the Efficient Market Hypothesis, and to study how correlations between assets change over time, especially during crises.

- Use Case: A study might analyse the correlation between Bitcoin and the S&P 500 (as done in this EDA) to determine if cryptocurrency acts as a "safe haven" asset or if it is increasingly correlated with the broader stock market.
- Example Publication: Papers in journals like the *Journal of Finance* or the *Journal of Financial Economics* often use similar datasets to model volatility and asset correlation dynamics. A notable example is the study by Baur, D.G., & Dimpfl, T. (2021) titled "The volatility of Bitcoin and its role as a safe haven and an diversifier," which explores these very relationships.
- Link: <https://link.springer.com/article/10.1007/s00181-020-01990-5>

2. Financial Institutions' Risk Management and Portfolio Allocation

Investment banks, hedge funds, and asset managers use this type of data to build risk models and optimize portfolio allocation. The statistical measures calculated in this EDA—such as volatility (standard deviation), skewness, and kurtosis—are critical inputs for these models.

- Use Case: A portfolio manager would use the correlation matrix to construct a diversified portfolio, combining assets with low correlation to reduce overall risk. The "fat tails" (high kurtosis) identified in the returns distribution are essential for Value-at-Risk (VaR) calculations, which estimate potential losses.

- Example Publication: Major financial institutions like Goldman Sachs, J.P. Morgan, and BlackRock regularly publish market analysis reports for their clients that are based on the EDA of similar multi-asset datasets.

3. Financial Journalism and Economic Reporting

Leading financial news outlets use this data to create the charts and analyses that inform the public about market trends. The visualizations in this report, such as the normalized sector performance charts, are staples of financial journalism.

- Use Case: A journalist at Bloomberg or The Wall Street Journal would create a normalized performance chart to write a story about how the technology sector has outperformed the broader market since the pandemic.
- Example: The data visualization teams at outlets like the Financial Times and The Economist are renowned for their use of such datasets to explain complex economic events to the public. Their work is a real-world application of the EDA process conducted in this project.

CHAPTER 9

CONCLUSION AND APPENDIX

This Exploratory Data Analysis has successfully navigated the complexities of a diverse, five-year financial dataset, transforming raw market data into a coherent narrative of trends, risks, and opportunities. Through a methodical process of data cleaning, preprocessing, and extensive visualization, this report illuminates the profound dynamics of the U.S. stock market and global commodities. The initial transformation phase was critical; standardizing column structures, correcting data types, and imputing missing values established a reliable foundation for the analysis.

The analysis conclusively demonstrates that the market's trajectory was not monolithic. It was characterized by the aggressive, technology-driven growth of the Nasdaq 100, which outpaced the broader market, and the extreme volatility of the cryptocurrency sector. In contrast, traditional assets like Gold fulfilled their role as stable safe havens, while commodities like Copper and Crude Oil followed more cyclical patterns. The high correlation observed within sectors—particularly among tech stocks and cryptocurrencies—underscores the importance of diversification.

Furthermore, this EDA has successfully prepared the dataset for the next logical phase of the project. The identified patterns and correlations provide a solid foundation for building robust predictive models. In essence, this analysis provides a clear, evidence-based picture of recent market history while paving the way for developing sophisticated tools to anticipate future market behavior.

APPENDIX

Dataset Name: Stock Market Dataset.csv

Dataset Link: <https://www.kaggle.com/datasets/saketk511/2019-2024-us-stock-market-data>

GitHub Link:

https://github.com/PrathamAgrawal51/Pratham_Agrawal_22070521078_DS_CA1

Name: Pratham Agrawal

PRN:22070521078

Sem: 7th

Sec: C