CSCIGA 2437

Analysis of Cryptocurrency and Industry Stocks

Yi Peng
Chenmeinian Guo
Xiao Lu
Weizhou Ding

Overview

- Abstract & Motivation
- Information Technology
- Real Estate
- Crypto
- Energy
- Summary

NYU

Abstract

The purpose of this project is to use some data analysis methods to analyze historical data of stocks, cryptocurrencies, and real estate, study their relationships, and give some investment suggestions

Motivation

Who will benefit from this analytic?

Investors, financial analysts, and stock managers

Why is this analytic important?

This analytic helps investors and analysts find links between cryptocurrencies and stocks, making it easier to manage risks, and boost returns. It’s all about giving clear insights for smarter investment decisions in today’s fast-moving markets.

Using clustering to get correlations

Data ingestion	Data cleansing	Normalization	Clustering
Extract stocks whose stock sector is technology information and concentrate them into the same dataframe	Extract stock data from 2015 to 2019, remove outliers, and replace null values with 0	Considering the price differences between stocks, regularize the stock data and use the rate of change to represent the trend of the stock.	Use k-means to cluster stocks and divide them into 5 categories, using the rate of change as the dimension

kaggle

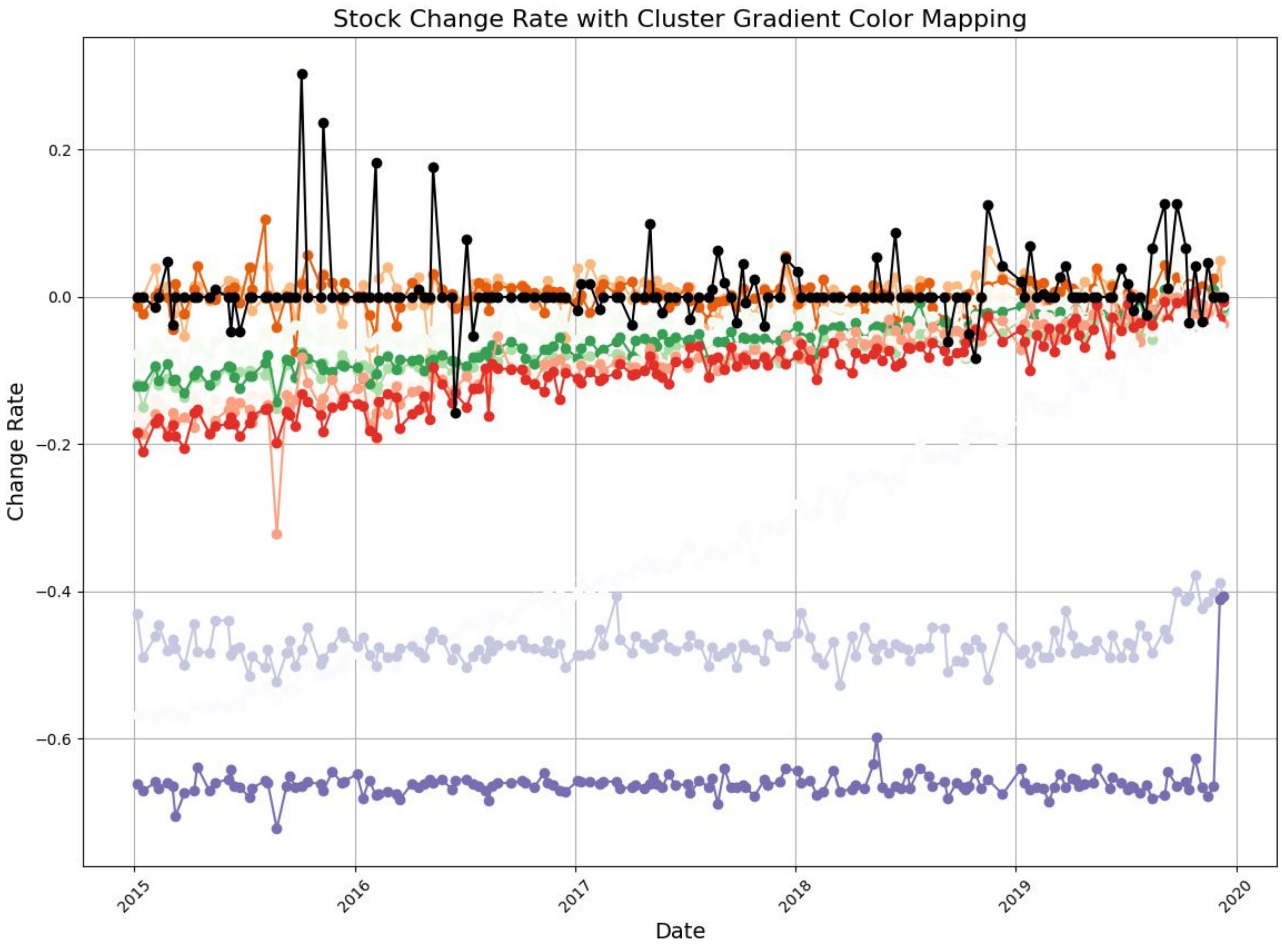
AMEX, NYSE, NASDAQ stock histories

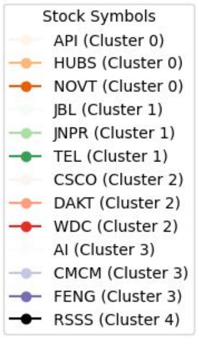
End-of-day data of 8,000+ stocks trading on AMEX, NYSE, and NASDAQ

“Information Technology” Stock

stock symbols	2015-01-05	2015-01-15	2015-02-05	2015-02-10
HUBS	-0.00149429881735...	6.152094369817895E-4	0.03959245038616732	0.010465022665288926
NOVT	-0.01187151396067343	-0.02252909988809...	7.391152312787863E-4	0.005860800435294...
API	0.0	0.0	0.0	0.0
SONM	0.0	0.0	0.0	0.0
MTCH	0.0	0.0	0.0	0.0
CMCM	-0.4307012370955561	-0.4899451425547363	-0.46096880561830855	-0.44568279457790083
JBL	-0.07678201402293439	-0.09697739634400548	-0.06186853074330...	-0.06854709359102473
TTWO	-0.03662362170444322	-0.00431340521930...	-0.01808869601015206	-0.02661990689466...
CSCO	-0.16201027845894045	-0.17443983092566506	-0.13782864864741529	-0.1436791828345364
MX	-0.01639344506385...	-0.02677883978329...	0.005280523149862271	-0.00665332199414...
GIB	0.009157019736254712	0.009566850920357695	0.013118781170277316	0.013922271559746902
GEN	-0.04952820109573586	0.02339179003549785	0.03629981510555358	0.009999990463256836
TEL	-0.12098070049409004	-0.12105467337614684	-0.09416167805818954	-0.11269070477000276

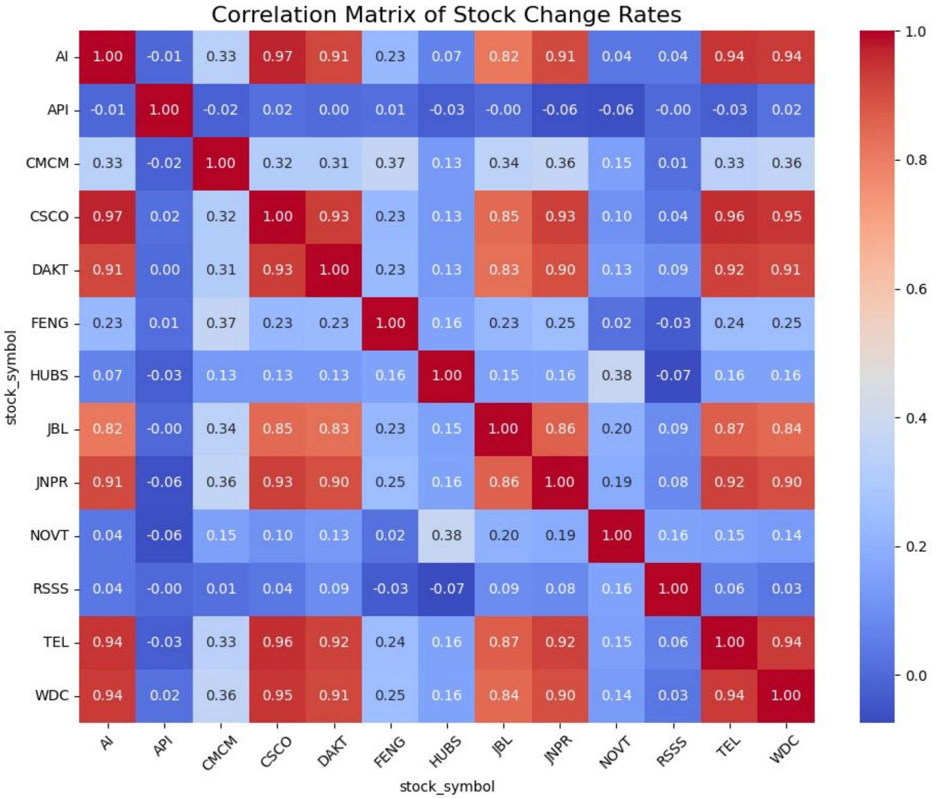
Each cell represents the stock's rate of change for the day





stock_symbol	prediction
JBL	1
TEL	1
JNPR	1
CMCM	3
FENG	3
AI	3
RSSS	4
CSCO	2
DAKT	2
WDC	2
HUBS	0
NOVT	0
API	0

Correlations



- Get 5 clusters
- For every cluster, select 3 stocks
- Calculate the correlation

stock_symbol	prediction
JBL	1
TEL	1
JNPR	1
CMCM	3
FENG	3
AI	3
RSSS	4
CSCO	2
DAKT	2
WDC	2
HUBS	0
NOVT	0
API	0

Real Estate

Dataset

Publicly available NYC property sales dataset from Kaggle, containing property sale records:

BOROUGH, NEIGHBORHOOD, CATEGORY, SALE PRICE, SALE DATE, YEAR BUILT, RESIDENTIAL UNITS, TOTAL UNITS, etc.

The **focus** is on columns like SALE PRICE, CATEGORY, NEIGHBORHOOD, and SALE DATE for analysis.

Overview

Data Processing

Analysis:

- **Zero Sale**
- **Category, Neighborhood, Zip Code**
- **Time (Month & Quarter)**

Clustering

Current NYC Property Sales

Dataset · 8mo ago · by Data Science Donut

15+ year's worth of sold NYC real estate properties. Updated monthly.

Data Processing

Steps for Cleaning the Data

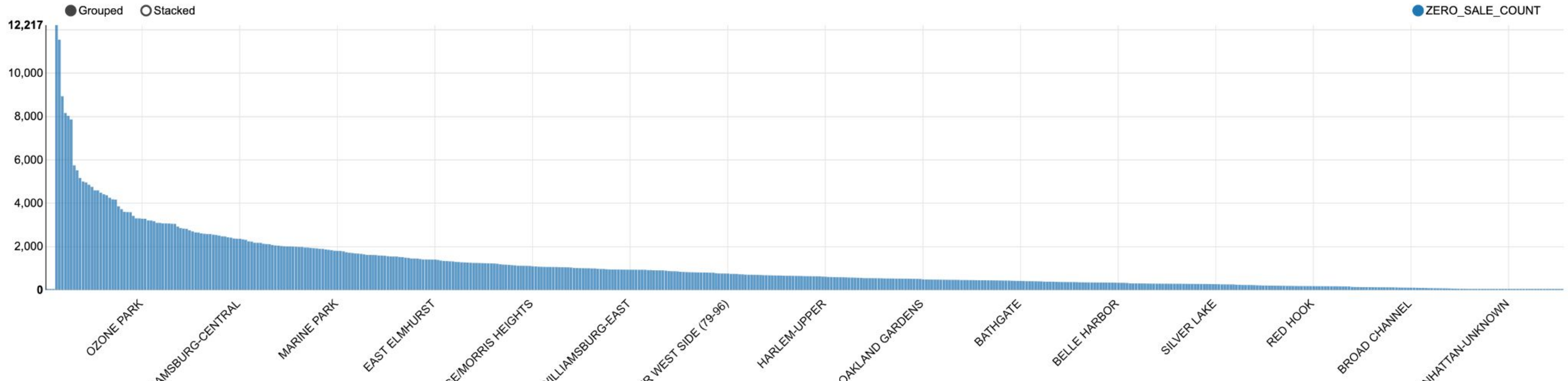
1. **Null/Invalid Value Handling**: Removed rows with critical null or just invalid values in columns such as `SALE PRICE` and `ZIP CODE`.
2. **Column Filtering**: Dropped irrelevant columns such as `EASE-MENT` and `APARTMENT NUMBER`.
3. **Outlier Removal**: Removed extreme values in columns like `SALE PRICE`, `TOTAL SALES`, and `GROWTH RATE` using 1st and 99th percentiles.

Note: An easement is a nonpossessory right that allows another party to use your land for specific purposes without owning the land itself.

Zero Sale Analysis

About **27.8% of the records have `SALE PRICE = 0`**. These often indicate non-standard transactions like inheritances, donations, family transfers, or government acquisitions. For example, neighborhoods like **Midtown West** (12,217) and **Flushing-North** (11,537) lead in a lot of zero-sale counts, potentially reflecting high activity in corporate or institutional transfers. In contrast, areas like **Bedford Stuyvesant** (8,930) and **Crown Heights** (5,160) might reflect more localized, community-driven property exchanges, such as inheritances or family transfers.

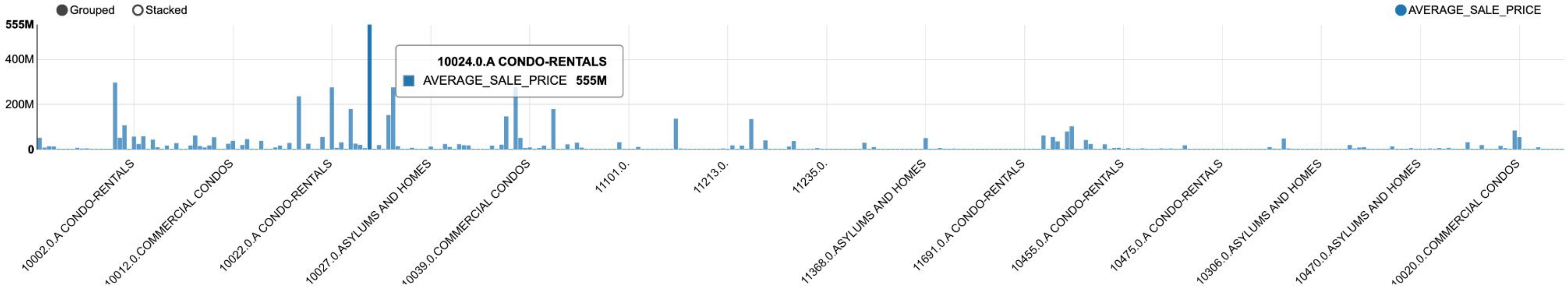
Understanding these patterns helps identify where such non-market activities are concentrated, providing insights into urban property trends and socio-economic behaviors within different neighborhoods.



Category Analysis

Some Findings from Category Analysis:

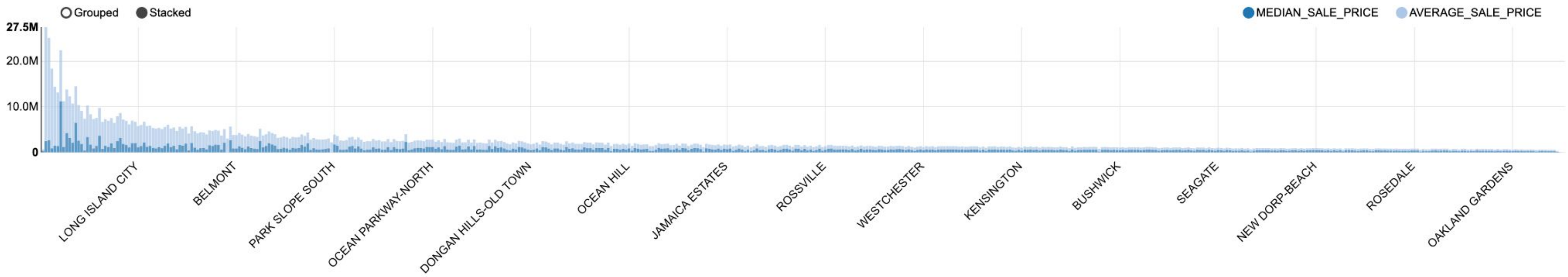
- **Growth Rate**: `Government` related categories exhibited unusually high growth rates due to specific outliers.
- **High-Value Categories**: `Theatres` and `Office Buildings` had the highest average and median sale prices.
- **Sales Volume**: Residential categories (i.e. `Coops Elevator Apartments`, `One Family Dwellings`) had the highest total sales.



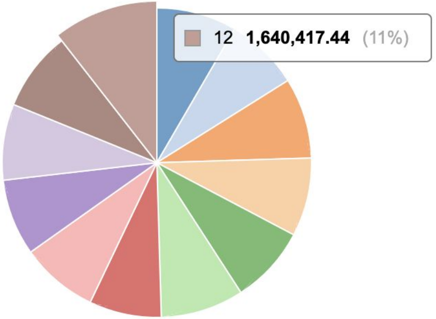
Neighborhood Analysis

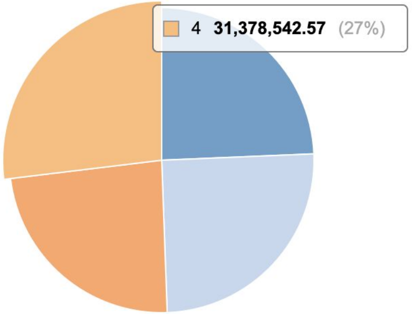
Some Findings from Neighborhood Analysis:

- **High-Value Neighborhoods**: Premium areas like `Midtown West` and `Wall Street` consistently had high average sale prices.
- **Localized Patterns**: Neighborhoods like `Bedford Stuyvesant` reflected community-driven transactions such as family transfers.



Time-Based Analysis





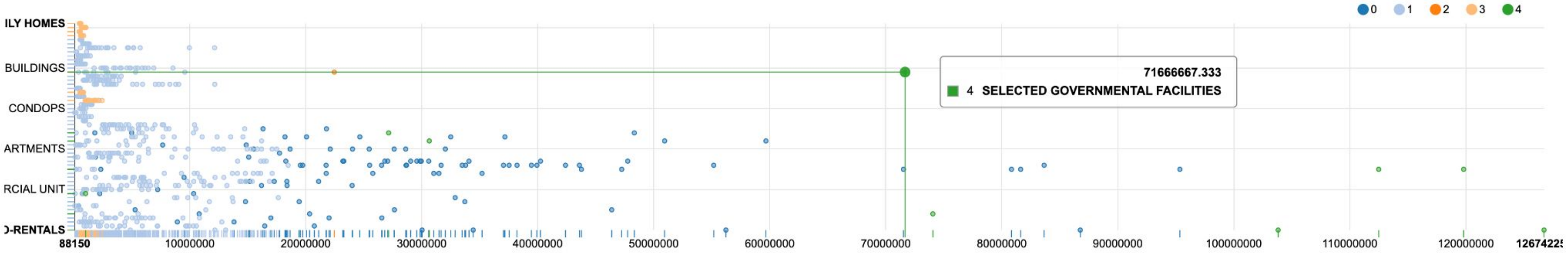
Key Insights from Temporal Analysis:

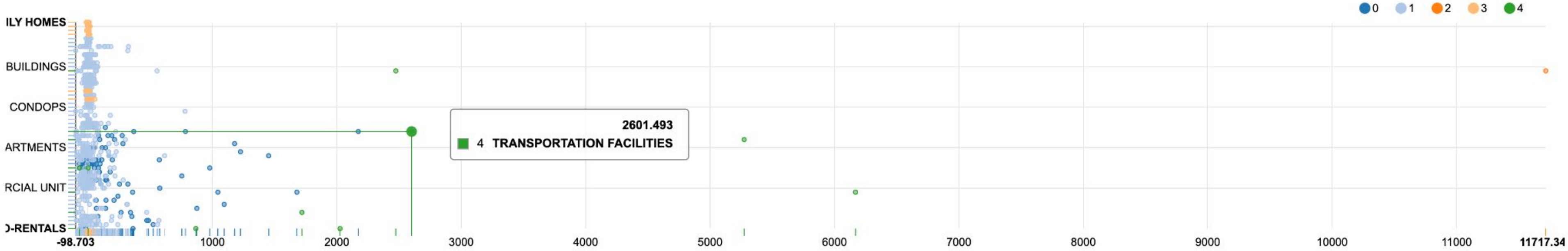
- **Monthly Trends**:
  - December had the highest average sale prices, while June and August had the highest total sales.
- **Quarterly Trends**:
  - Q4 slightly outperformed other quarters in average sale prices.
  - Q3 had the highest total sales volume, which potentially aligns with market activity in summer months.

NYU

Clustering

- Grouped properties into **five clusters** based on `SALE PRICE`, `GROWTH RATE`, and `TOTAL SALES` using **K Means** Algorithm.
- Identified **high-growth clusters** with categories and neighborhoods showing promising market potential.
- PCA-based visualization highlighted clear segmentation.
- X-axis for the below two graphs are average sale price and growth rate, respectively.



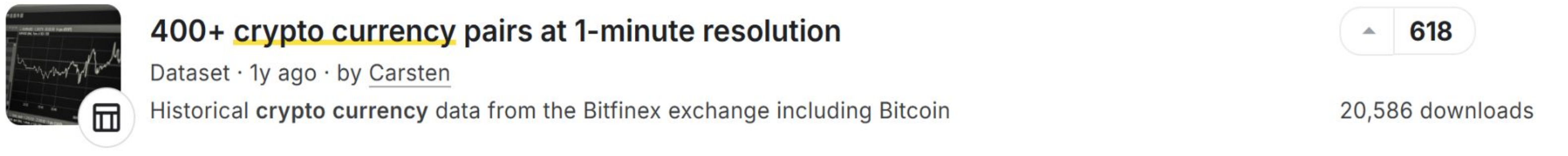


Data Source

Name: Crypto Currency pairs at 1-minute resolution

Description: This dataset provides open, high, low, close data at 1 minute resolution of BTC and ETH from 2013 to 2023.

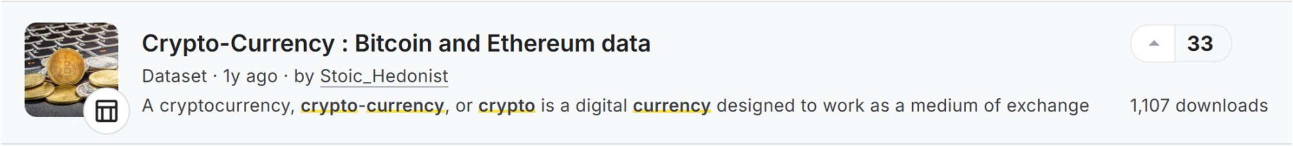
Size of data: 2.2 GB



Name: Crypto-Currency : Bitcoin and Ethereum data

Description: This data provides daily BTC and ETH transaction data from 2018 through 2023

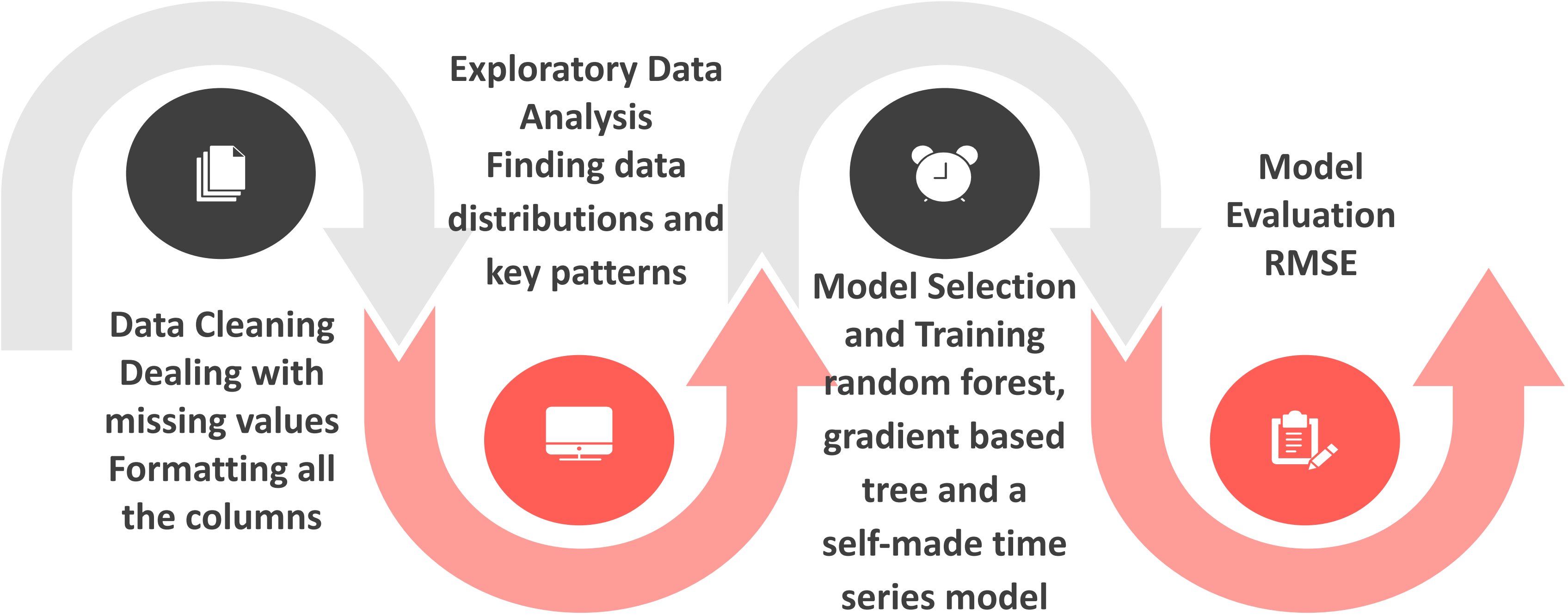
Size of data: 300 KB



Data Sample

12/28/2018	116.898201	137.647018	115.69313	137.647018	137.647018	3.130201009E9	Etherium
12/29/2018	138.468781	147.034332	134.570175	138.018341	138.018341	3.169029972E9	Etherium
12/30/2018	137.627457	140.689087	133.98233	139.859451	139.859451	2.660086834E9	Etherium
12/31/2018	140.031067	140.181152	132.519394	133.368256	133.368256	2.358360234E9	Etherium
1/1/2019	133.418152	141.397507	132.650711	140.819412	140.819412	2.258709868E9	Etherium
1/2/2019	141.519516	156.929138	140.650955	155.047684	155.047684	3.328240369E9	Etherium
1/3/2019	155.196045	155.863052	147.198364	149.13501	149.13501	2.67616488E9	Etherium
1/4/2019	148.912888	156.878983	147.907104	154.58194	154.58194	3.126192535E9	Etherium
1/5/2019	154.337418	160.82489	154.337418	155.638596	155.638596	3.33821192E9	Etherium
1/6/2019	155.80423	159.371445	152.085922	157.746201	157.746201	3.231294371E9	Etherium
1/7/2019	157.809494	158.450424	151.150726	151.699219	151.699219	2.712108388E9	Etherium
1/8/2019	151.967545	153.625778	148.66954	150.359634	150.359634	2.45980814E9	Etherium
1/9/2019	150.554688	153.622253	150.288376	150.803116	150.803116	2.369241197E9	Etherium
1/10/2019	150.843506	152.14827	126.529373	128.625183	128.625183	3.397734456E9	Etherium
1/11/2019	127.813965	130.165939	125.244942	127.548325	127.548325	2.667585234E9	Etherium
1/12/2019	127.528084	128.666122	125.446754	125.96653	125.96653	2.21210922E9	Etherium
1/13/2019	125.907227	126.267876	116.085968	116.897804	116.897804	2.268263944E9	Etherium

Design Diagram



NYU

Code Challenge 1

Standardizing the date formats across datasets

```
+-------------+
|    timestamp|
+-------------+
|1364774820000|
|1364774880000|
|1364774940000|
|1364775060000|
|1364775120000|
|1364775180000|
|1364775960000|
|1364776080000|
|1364776500000|
|1364777580000|
|1364777640000|
|1364778000000|
```

Use from_unixtime to format

```
| 12/1/2018|113.397758|120.841454|111.619125|
| 12/2/2018| 118.26815| 120.56205|116.092041|
| 12/3/2018|116.378761|116.619064|107.415657|
| 12/4/2018|108.803162|113.142914|107.402718|
| 12/5/2018|110.335518| 110.59861|102.475555|
| 12/6/2018|102.450592|104.103493| 91.761055|
| 12/7/2018| 91.649834| 96.089844| 83.469719|
| 12/8/2018|  93.41008| 97.059883| 86.825668|
| 12/9/2018| 92.035431| 98.900032| 91.031281|
```

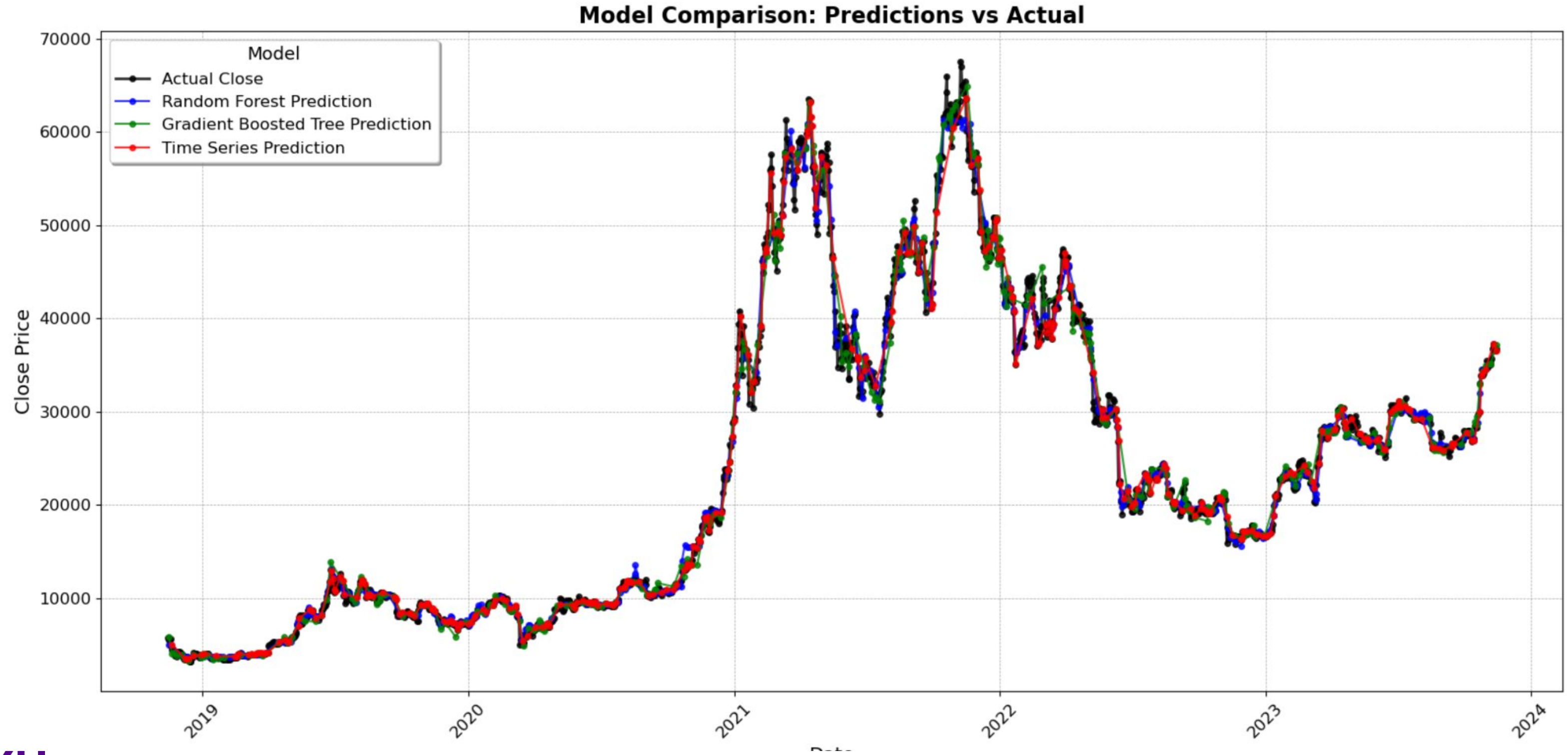
```
|  1/1/2019|133.418152|141.397507|132.650711|
|  1/2/2019|141.519516|156.929138|140.650955|
|  1/3/2019|155.196045|155.863052|147.198364|
|  1/4/2019|148.912888|156.878983|147.907104|
|  1/5/2019|154.337418| 160.82489|154.337418|
|  1/6/2019| 155.80423|159.371445|152.085922|
|  1/7/2019|157.809494|158.450424|151.150726|
|  1/8/2019|151.967545|153.625778| 148.66954|
|  1/9/2019|150.554688|153.622253|150.288376|
| 1/10/2019|150.843506| 152.14827|126.529373|
| 1/11/2019|127.813965|130.165939|125.244942|
```

Use regular expressions

Code Challenge 2

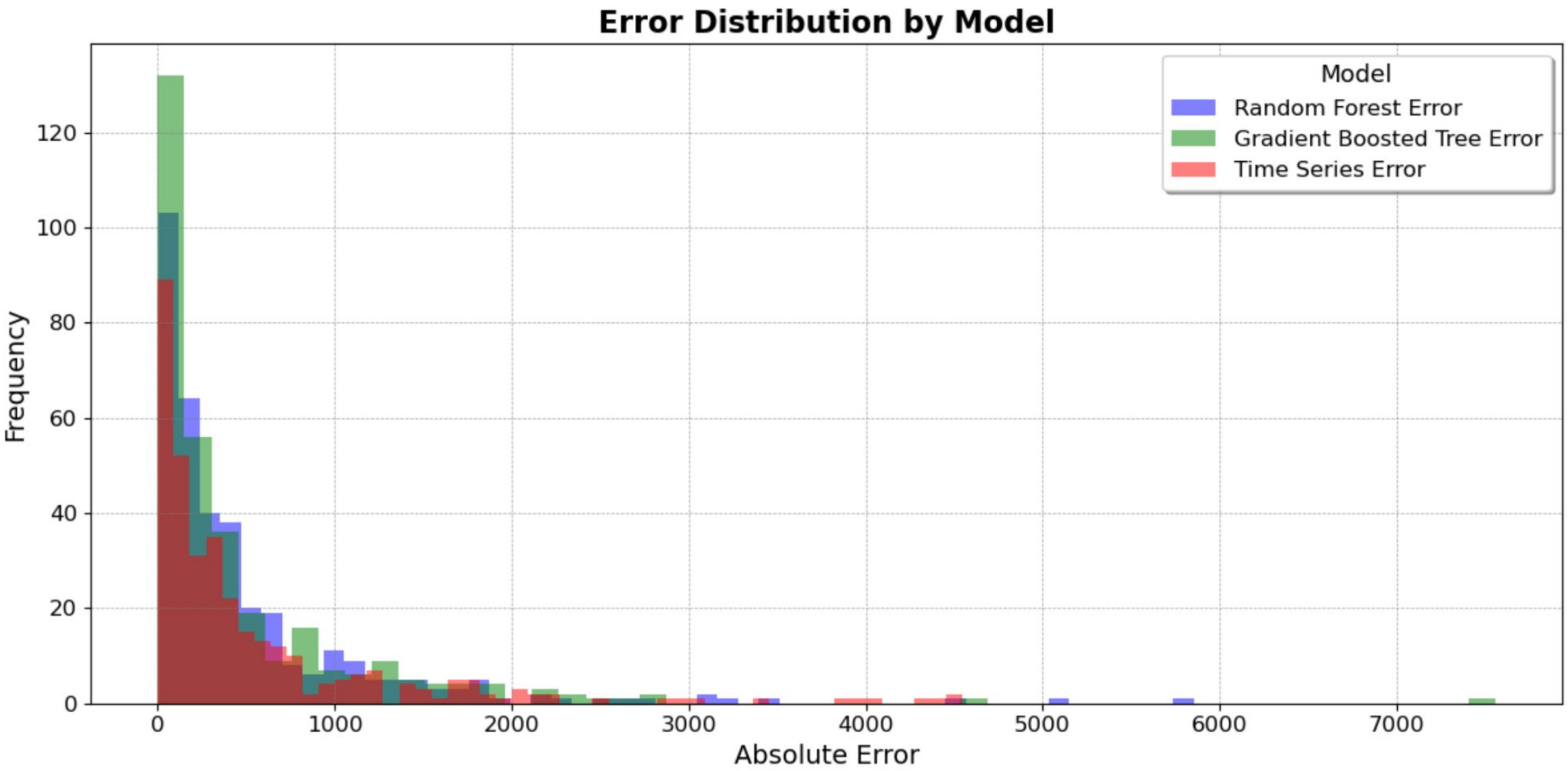
- Regression Task and nonlinear data
- Random Forest and Gradient Boosted Trees
- ARIMA
- Inconsistency with the current Scala versions
- A simple time series model using Linear Regression combined with lagged features

Result



Model Comparison: Predictions vs Actual
Model
Actual Close
Random Forest Prediction
Gradient Boosted Tree Prediction
Time Series Prediction
Close Price
70000
60000
50000
40000
30000
20000
10000
2019
2020
2021
2022
2023
2024

Result



Error Distribution by Model
Model
Random Forest Error
Gradient Boosted Tree Error
Time Series Error
Frequency
Absolute Error
0
20
40
60
80
100
120
0
1000
2000
3000
4000
5000
6000
7000

Energy Stocks

Data Source

Ticker	Company Name	Primary Industry	Secondary Industry	Tertiary Industry	Quaternary Industry
A.N	Agilent Technologies, Inc.	Health Care	Pharmaceuticals, Biotechnology & Life Sciences	Life Sciences Tools & Services III	Life Sciences Tools & Services
AA.N	Alcoa Corporation	Materials	Materials II	Metals & Mining	Aluminum
AACG.O	ATA Creativity Global	Consumer Discretionary	Consumer Services	Diversified Consumer Services	Education Services
AACT.N	Ares Acquisition Corp II	Financials	Diversified Financials	Capital Markets	Investment Banking & Brokerage
AACT_U.N	Ares Acquisition Corp II	Financials	Diversified Financials	Capital Markets	Investment Banking & Brokerage
AADI.O	Aadi Bioscience, Inc.	Health Care	Pharmaceuticals, Biotechnology & Life Sciences	Pharmaceuticals III	Pharmaceuticals
AAL.O	American Airlines Group Inc.	Industrials	Transport	Airlines III	Airlines
AAM.N	AA Mission Acquisition Corp.	Financials	Diversified Financials	Capital Markets	Investment Banking & Brokerage
AAME.O	Atlantic American Corporation	Financials	Insurance II	Insurance III	Multi-line Insurance
AAMU.N	AA Mission Acquisition Corp.	Financials	Diversified Financials	Capital Markets	Investment Banking & Brokerage
AAOI.O	Applied Optoelectronics, Inc.	Information Technology	Technology Hardware And Device	Communications Equipment III	Communications Equipment
AAON.O	AAON, Inc.	Industrials	Capital Goods	Building Products III	Building Products
AAP.N	Advance Auto Parts, Inc.	Consumer Discretionary	Retailing	Specialty Retail	Automotive Retail
AAPL.O	Apple Inc.	Information Technology	Technology Hardware And Device	Computers & Peripherals	Computer Hardware
AAT.N	American Assets Trust, Inc.	Real Estate	Real Estate II	Equity Real Estate Investment Trusts (REITs)	Diversified REITs
AB.N	Alliancebernstein Holding L.P.	Financials	Diversified Financials	Capital Markets	Asset Management & Custody Banks
ABAT.O	American Battery Technology Com	Materials	Materials II	Metals & Mining	Diversified Metals & Mining
ABBV.N	Abbvie Inc.	Health Care	Pharmaceuticals, Biotechnology & Life Sciences	Pharmaceuticals III	Pharmaceuticals
ABCB.N	Ameris Bancorp	Financials	Banks	Commercial Banks	Regional Banks
ABCL.O	Abcellera Biologics Inc.	Health Care	Pharmaceuticals, Biotechnology & Life Sciences	Biotechnology III	Biotechnology
ABEO.O	Abeona Therapeutics Inc.	Health Care	Pharmaceuticals, Biotechnology & Life Sciences	Pharmaceuticals III	Pharmaceuticals

NYU

Energy Stocks

Data Cleaning and Manipulation

- Selecting stocks
- Cleaning the data format
- Downloading the actual price data for later analysis

NYU

Energy Stocks

Date in use

	Date	ACDC	AE	ALTO	AM	AMPY	AMR	AMTX	\
0	2022-05-13	18.110001	32.203209	4.86	8.433784	6.98	151.680298	7.66	
1	2022-05-16	17.990000	31.971334	4.81	8.516552	7.33	143.772537	7.35	
2	2022-05-17	18.080000	31.850763	4.88	8.789676	8.00	147.379929	8.21	
3	2022-05-18	17.980000	32.110462	4.48	8.533106	7.70	144.987671	7.80	
4	2022-05-19	16.879999	32.397991	4.33	8.351019	7.49	143.639633	7.63	

	APA	AR	...	WFRD	WHD	WKC	WMB	\
0	38.455154	32.279999	...	29.977989	46.279037	22.015083	30.633873	
1	39.632633	33.939999	...	30.942776	47.567005	22.807402	31.206215	
2	40.314827	36.270000	...	31.290894	47.586514	22.581028	31.611265	
3	37.800991	34.799999	...	29.600033	45.937531	24.090200	31.021301	
4	38.146755	35.169998	...	29.629871	46.679092	24.241116	30.836391	

	WTI	WTTR	WWR	XOM	XPRO	YPF
0	5.397073	7.200490	1.17	81.635147	12.42	4.25
1	5.612565	7.539779	1.14	83.555229	12.82	4.42

Energy Stocks

Daily Return

Date	ACDC_ret	AE_ret	ALTO_ret	AM_ret	AMPY_ret	AMR_
2022/7/21	-0.017292839	3.176991803490	-0.081967191	-0.015151446	-0.051724125	-0.
2022/7/22	-0.049150454	-0.006986319	-0.045918383	-0.010256383	-0.039669461	-
2022/7/25	0.068283324	-0.001279266	0.032085530816	0.031087998451	0.11703955852	0.0
2022/7/26	0.022102798	0.015690111594	-0.005181342	0.003598767	-0.010785778	-
2022/7/27	0.062536508	-0.004413685	0.044270916	0.009221514	0.073208689	0.0
2022/7/28	0.004950503	0.010766338601	0.004987526	0	-0.029027549	-0.0
2022/7/29	0.006020728	0.065162915	0.06699751	0.021319841	0.02391627	0.011
2022/8/1	-0.02285093	-0.0117647	-0.020930267	-0.001988214	-0.048175172	-0.0
2022/8/2	0.022828607	-0.005952408	0.06175766	9.96199724079	-0.003067482	4.430
2022/8/3	-5.44E-04	0.019760589261	-0.008948538	-0.002985289	-0.049230796	-0.0
2022/8/4	-0.065359517	-0.01673515	-0.018058674	-0.038922033	-0.03398059	-0.0
2022/8/5	0.0180652371036	0.041803468	0.082758653	0.00934553	0.008375242	0.0
2022/8/8	0.002862115	-0.040412708	0.023354593	0.001029112	-0.013289024	0.0
2022/8/9	-0.027397233	5.972583358638	-0.01244812	0.01438836430	0.0269360009	-0.0
2022/8/10	0.047535177	0.069552207	0.037815088168	0.02127664187	0.027868865	0.0
2022/8/11	0.026330493	-0.020373807	-0.008097158	0.029761808	0.1355661734	0.0
2022/8/12	0.1080785794333	-0.038461494	0.028571400767	0.004817085	0.05617979	0.0
2022/8/15	-0.033497459	-0.094222455	-0.053571425	-0.016299228	-0.019946821	-0.0

Energy Stocks

Annual Return and Volatility

```scala
// Annual return = (1 + daliy return)^(252) - 1
// Annudal volatity = daily std * sqrt(252)

val annualFactor = math.sqrt(252)

// avg return and stf
val meanAggExprs = retCols.map(c => mean(c).alias(c+"_mean"))
val stdAggExprs = retCols.map(c => stddev(c).alias(c+"_std"))

val aggExprs = retCols.flatMap(c => Seq(mean(c).alias(c+"_mean"), stddev(c).alias(c+"_std")))
val statsRow = dfRet.agg(aggExprs.head, aggExprs.tail:_*).first()

// (Stock, AnnualReturn, AnnualVol)
val stats = stockCols.map { s =>
  val meanVal = statsRow.getAs[Double](s+"_ret_mean")
  val stdVal = statsRow.getAs[Double](s+"_ret_std")
  val annualRet = math.pow(1.0 + meanVal, 252) - 1.0
  val annualVol = stdVal * annualFactor
  (s, annualRet, annualVol)
}
```

```
statsDF.show()

+-----+--------------------+------------------+
|Stock|        AnnualReturn|         AnnualVol|
+-----+--------------------+------------------+
| ACDC|-0.21029499005275976| 0.639152612416709б|
|   AE|0.048701821034189186|0.40399807776080177|
| ALTO| -0.0316155464543727| 0.7773455073746321|
|   AM| 0.34039834112930256|0.23514694745993284|
| AMPY| 0.19982116975682107| 0.5677459298670482|
|  AMR|  0.3669534728158701| 0.5215014775327921|
| AMTX|  0.1532179755393679| 0.9945778209173148|
|  APA|-0.05739857075947874| 0.4158401390193946|
|   AR|-0.01480804956148...| 0.4483932489475088|
| ARCH| 0.19110044084595024| 0.415588903296663 5|
| AREC| 0.06336099503454085| 0.8340174702447463|
| ARLP|   0.281631916616516|  0.30278382619460 6|
| AROC|  0.6588223687936339|0.34452030919017446|
|  ASC| 0.46834482159030677| 0.4302737789184405|
```

Energy Stocks

Return and Volatility Threshold

```
val percentileAnnualRet = 0.8 // top 20% return
val percentileAnnualVol = 0.5 // top 50% low vol

def percentile(arr: Array[Double], p: Double): Double = {
  val sortedArr = arr.sorted
  val idx = ((sortedArr.length - 1) * p).toInt
  sortedArr(idx)
}

val returnsArray = stats.map(_._2).toArray
val volsArray = stats.map(_._3).toArray

val retThreshold = percentile(returnsArray, 1 - percentileAnnualRet)
val volThreshold = percentile(volsArray, percentileAnnualVol)

val selected = statsDF.filter(col("AnnualReturn") > retThreshold && col("AnnualVol") < volThreshold)
selected.show(false)
```

NYU

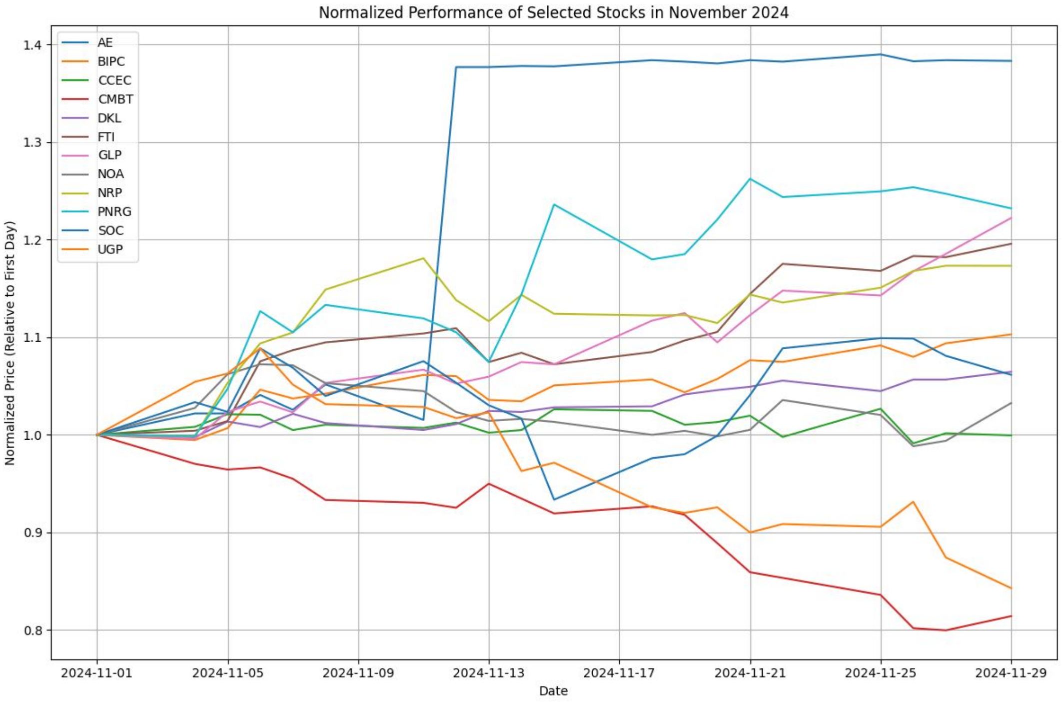
Energy Stocks

Portfolio Management

- Sort high return and low volatility
- Adding them in order if cor < 0.3

```scala
val filteredStocks = statsArr.filter { case (s, annRet, annVol) =>
  annRet > retThreshold && annVol < volThreshold
}

val sortedFiltered = filteredStocks.sortBy(_._2)(Ordering[Double].reverse)
val stockIndexMap = stockCols.zipWithIndex.toMap
val corrThreshold = 0.3

var selectedStocks = Seq[String]()

for ((stk, ret, vol) <- sortedFiltered) {

  if (stockIndexMap.contains(stk)) {
    val idx = stockIndexMap(stk)
    val isLowCorr = selectedStocks.forall { chosen =>
      val chosenIdx = stockIndexMap(chosen)
      val corrVal = corrMatrix(idx, chosenIdx)
      math.abs(corrVal) < corrThreshold
    }
    if (isLowCorr) {
      selectedStocks = selectedStocks :+ stk
    }
  }
}
```

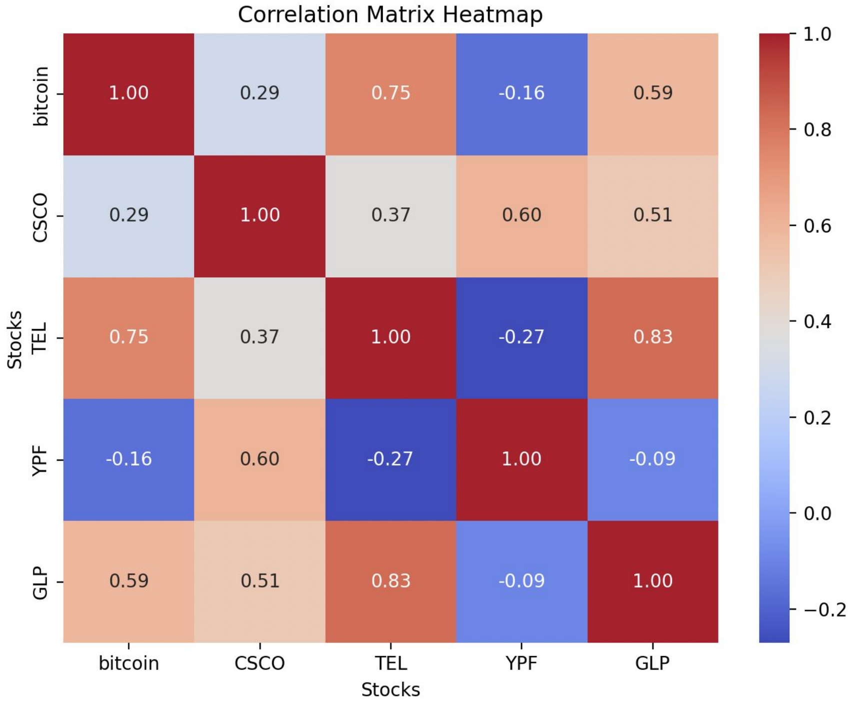
Results



SelectedStock
FTI
NRP
SOC
GLP
PNRG
CMBT
UGP
NOA
CCEC
BIPC
AE
DKL

Conclusions

- Cisco and Western Digital also have a high positive correlation
- Bitcoin and TEL stock prices exhibit high positive correlation
- Consider Bitcoin for long-term diversification, but stay cautious due to its high volatility and market risks.



NYU

Acknowledgements & References

Jiun Yen(2018)Stock diversity analysis 1. Kaggle.

https://www.kaggle.com/code/qks1lver/stock-diversity-analysis-1

Stoic_Hedonist（2020）Time series analysis : LSTM - Bitcoin, Etherium 1.Kaggle

https://www.kaggle.com/code/zsinghrahulk/time-series-analysis-lstm-bitcoin-etherium

NYU


NYU
Thank you