**School of Applied Computer Science & Information Technology**

**SENG8081 - Case Studies**

# **Cryptocurrency Market Analysis and Forecasting: Comparative Study of Bitcoin and Ethereum Prices (2018–Present)**

**Team Members:**

**Krishna Vamsi Cherukupalli**

**Lohith Reddy Danda**

**Paras Rupani**

**Instructor: Prof. David Marsh**

**Date: July 31, 2025**

Table of Contents

[**Cryptocurrency Market Analysis and Forecasting: Comparative Study of Bitcoin and Ethereum Prices (2018–Present)** 1](#_Toc204898367)

[**Abstract** 4](#_Toc204898368)

[**1.** **Introduction** 5](#_Toc204898369)

[**2.** **Data Research and Integration** 5](#_Toc204898370)

[**Data Sources and Acquisition:** 5](#_Toc204898371)

[**Integration and Standardization:** 5](#_Toc204898372)

[**Unified Datasets:** 6](#_Toc204898373)

[**Reflection:** 6](#_Toc204898374)

[**3.** **Data Collection** 6](#_Toc204898375)

[**Automated Data Extraction:** 6](#_Toc204898376)

[**Scope and Scale:** 6](#_Toc204898377)

[**Data Storage:** 6](#_Toc204898378)

[**Challenges and Solutions:** 6](#_Toc204898379)

[**Outcome:** 7](#_Toc204898380)

[**4.** **Data Storage and Maintenance** 7](#_Toc204898381)

[**Initial Storage Approach:** 7](#_Toc204898382)

[**Version Control and Collaboration:** 7](#_Toc204898383)

[**Data Maintenance Practices:** 7](#_Toc204898384)

[**Scalability Considerations:** 7](#_Toc204898385)

[**Reflection:** 8](#_Toc204898386)

[**5.** **Data Quality** 8](#_Toc204898387)

[**Defining Data Quality for Crypto Markets:** 8](#_Toc204898388)

[**Quality Assurance Workflow:** 8](#_Toc204898389)

[**Why This Matters:** 8](#_Toc204898390)

[**Reflection:** 9](#_Toc204898391)

[**6.** **Data Analysis and Visualization** 9](#_Toc204898392)

[**Exploratory Data Analysis (EDA):** 9](#_Toc204898393)

[**Comparative Analysis: Bitcoin vs. Ethereum:** 9](#_Toc204898394)

[**Advanced Visualizations:** 9](#_Toc204898395)

[**Forecasting and Modeling:** 9](#_Toc204898396)

[**Key Insights:** 10](#_Toc204898397)

[**Reflection:** 10](#_Toc204898398)

[**Prices of Bitcoin & Ethereum over the year** 10](#_Toc204898399)

[**Return Percentage Over the Years by Horizon Label BTC/ETH and Coin** 11](#_Toc204898400)

[**Average Daily Trading Volume of Bitcoin vs 30 Day Moving Average** 11](#_Toc204898401)

[**Average Daily Trading Volume of Ethereum vs 30 Day Moving Average** 12](#_Toc204898402)

[**7.** **Extension** 12](#_Toc204898403)

[**Expanding to More Assets:** 13](#_Toc204898404)

[**Increasing Data Granularity and Real-Time Analytics:** 13](#_Toc204898405)

[**Scaling Data Storage and Processing:** 13](#_Toc204898406)

[**Real-Time Dashboards and Alerts:** 13](#_Toc204898407)

[**Reflection:** 13](#_Toc204898408)

[**8.** **Allocation Project Team Roles** 13](#_Toc204898409)

[**Team Member Roles and Contributions:** 13](#_Toc204898410)

[**Team Communication and Workflow:** 14](#_Toc204898411)

[**What Worked Well:** 14](#_Toc204898412)

[**What We would Improve Next Time:** 14](#_Toc204898413)

[**Reflection:** 14](#_Toc204898414)

[**9.** **Project Timeline** 14](#_Toc204898415)

[**Project Timeline Overview:** 15](#_Toc204898416)

[**How We Managed Our Timeline:** 15](#_Toc204898417)

[**Lessons Learned:** 15](#_Toc204898418)

[**Reflection:** 15](#_Toc204898419)

[**References** 16](#_Toc204898420)

# **Abstract**

The world of cryptocurrency is dynamic, unpredictable, and endlessly fascinating. In this project, we set out to explore the price movements of the two most prominent cryptocurrencies Bitcoin and Ethereum over the past seven years, from 2018 to the present. Our goal was not only to analyze their historical trends but also to forecast their future trajectories using modern data science techniques. By collecting and integrating data from multiple reputable sources, we compared the performance, volatility, and growth patterns of both assets. We also tackled the practical challenges of data quality, storage, and scalability, and reflected on how this project could be expanded in the future. Our findings offer valuable insights for investors, researchers, and anyone curious about the evolving landscape of digital currencies.

# **Introduction**

The cryptocurrency market is a fast-paced, data-rich environment, and Bitcoin and Ethereum remain at its core. Our project, Pipeline to Insights: Data Engineering and Visualization of Bitcoin Market Trends, is designed to go beyond daily price analysis. We focus on high-frequency, daily interval trading data for both Bitcoin and Ethereum, sourced from multiple exchanges. This approach allows us to capture the microstructure of the market revealing patterns, volatility, and trading behaviors that are invisible in daily data.

Our main objective is to build a robust data engineering pipeline that can ingest, clean, and process massive volumes of raw trading data, transforming it into actionable insights. We leverage Python for data extraction, cleaning, and transformation, and use advanced visualization techniques to make sense of the trends and anomalies in the world’s most prominent cryptocurrencies.

This project not only demonstrates technical proficiency in data engineering and analytics but also provides a practical framework for real-time market monitoring and research.

# **Data Research and Integration**

Our project’s foundation is built on the acquisition and integration of high-frequency, 1-hour interval trading data for both Bitcoin and Ethereum. This granular approach enables us to analyze market microstructure, spot short-term trends, and understand volatility at a much deeper level than traditional daily datasets.

**Data Sources and Acquisition:**

We sourced our data from multiple reputable exchanges and APIs, including Yahoo Finance and other open crypto data providers. The repository contains dedicated folders and scripts for data extraction, such as the YahooAPI\_BTC\_ETH and Data Extraction directories. These scripts automate the process of fetching raw trading data, ensuring that we capture every price movement, trade volume, and timestamp with precision.

## **Integration and Standardization:**

Integrating data from different exchanges and APIs presents unique challenges. Each source may have its own timestamp conventions, data formats, and even slight discrepancies in reported prices. To address this, we developed a robust ETL (Extract, Transform, Load) pipeline:

* **Timestamp Alignment:** The data is converted to a standard UTC format, ensuring consistency across sources.
* **Field Mapping:** We standardized key fields such as open, close, high, low, and volume, so that Bitcoin and Ethereum data can be compared directly.
* **Data Cleaning:** Scripts in the Data collection and cleaning codes and Data\_Cleaning\_Silver.py files handle missing values, outliers, and duplicate records, resulting in clean, analysis-ready datasets.

## **Unified Datasets:**

The output of our integration process is stored in btc\_cleaned.csv and eth\_cleaned.csv. These files represent unified, high-quality datasets that are ready for downstream analysis and visualization.

**Reflection:**

By investing in strong data research and integration process, we ensure that our insights are built on a solid, reliable foundation. This meticulous approach allows us to trust our results and confidently explore the nuances of crypto market behavior.

# **Data Collection**

Collecting high-frequency cryptocurrency data is a complex and resource-intensive task, but it is essential for understanding the true dynamics of the market. Our project’s data collection strategy is designed to capture every significant movement in the Bitcoin and Ethereum markets, down to the hour.

## **Automated Data Extraction:**

We developed a suite of Python scripts, found in the Data collection and cleaning codes, YahooAPI\_BTC\_ETH, and Data Extraction folders, to automate the process of gathering daily interval trading data. These scripts connect to APIs and data endpoints, systematically downloading price, volume, and timestamp information for both Bitcoin and Ethereum.

## **Scope and Scale:**

* **Timeframe:** Our dataset spans multiple years, providing a rich historical context for analysis.
* **Granularity:** By focusing on daily intervals, we capture the fine-grained fluctuations and rapid market responses that are often missed in daily data.
* **Assets:** While the primary focus is on Bitcoin, the same pipeline and methodology are applied to Ethereum, enabling direct comparison and cross-asset analysis.

**Data Storage:**

Raw data is initially stored in structured table within the repository. This format allows for easy inspection, sharing, and further processing. As the data volume grows, we have considered migrating to more scalable storage solutions, such as time-series databases, to support real-time analytics and larger datasets.

## **Challenges and Solutions:**

* **API Rate Limits:** We implemented intelligent throttling and retry mechanisms to avoid being blocked by data providers.
* **Missing or Corrupted Data:** Our scripts include validation steps to detect and handle missing or anomalous records, ensuring data integrity.
* **Data Consistency:** By running regular checks and cross-referencing multiple sources, we minimize discrepancies and maintain a high standard of data quality.

**Outcome:**

The result is a comprehensive, daily record of Bitcoin and Ethereum trading activity, ready for advanced analysis and visualization. This dataset forms the backbone of our project, enabling deep dives into market trends, volatility, and trading patterns.

# **Data Storage and Maintenance**

Handling high-frequency, high-volume cryptocurrency data requires thoughtful storage solutions and diligent maintenance practices. Our project’s data storage strategy is designed to ensure data integrity, accessibility, and scalability as our dataset grows.

## **Initial Storage Approach:**

We store our cleaned and processed data in table files specifically, btc\_cleaned.csv and eth\_cleaned.csv which are included in the repository. This format is both human-readable and compatible with Python’s data analysis libraries, making it easy to load, inspect, and share data among team members.

## **Version Control and Collaboration:**

All data files and scripts are managed through GitHub, providing robust version control. This ensures that every change whether it is a new data pull, a cleaning script update, or a bug fix is tracked and reversible. Team members can collaborate seamlessly, review each other’s work, and roll back to previous versions if needed.

## **Data Maintenance Practices:**

* **Regular Updates:** Our data extraction scripts are designed to run on a schedule, ensuring that our datasets remain current and comprehensive.
* **Validation and Backups:** After each data update, validation scripts check for missing values, duplicates, and anomalies. Cleaned data is backed up to prevent accidental loss.
* **Documentation:** Each data processing step is documented in the repository, making it easy for new contributors to understand the workflow and for the team to maintain transparency.

## **Scalability Considerations:**

As our dataset expands especially with daily interval data over multiple years CSV files may become unwieldy. We are exploring more scalable solutions, such as:

* **Databases:** Migrating to a relational database for faster queries and better data management.

**Reflection:**

By prioritizing good data storage and maintenance practices, we ensure that our analysis is always based on reliable, up-to-date information. This foundation allows us to scale our project and adapt to new research questions as the crypto market evolves.

# **Data Quality**

In a project that deals with high-frequency, high-volume trading data, maintaining data quality is both a challenge and a necessity. The accuracy of our insights and the reliability of our forecasts depend entirely on the integrity of our underlying data.

**Defining Data Quality for Crypto Markets:**

**For our project, data quality means:**

* **Completeness:** Every daily within our chosen timeframe should have a corresponding record for both Bitcoin and Ethereum.
* **Consistency:** All data fields (open, close, high, low, volume, timestamp) must be formatted uniformly and free from ambiguity.
* **Accuracy:** The data should reflect actual market activity, with no artificial spikes, gaps, or duplicated entries.
* **Reliability:** The data should be reproducible and verifiable, allowing others to replicate our analysis.

## **Quality Assurance Workflow:**

* **Automated Validation:** Our cleaning scripts (such as Data\_Cleaning\_Silver.py, Gold\_btc\_eth.py, and Platinum\_Btc\_Eth.py) include automated checks for missing values, duplicate timestamps, and out-of-range prices or volumes.
* **Cross-Source Verification:** When possible, we compare data from multiple sources (e.g., Yahoo Finance) to catch discrepancies. If inconsistencies are found, we investigate and resolve them before proceeding.
* **Handling Anomalies:** If a daily data is missing or appears anomalous (e.g., a sudden, unexplained price spike), our scripts flag it for review. We use interpolation or backup sources to fill small gaps, always documenting these interventions for transparency.
* **Changelog and Documentation:** Every correction or adjustment is logged in a changelog, ensuring that our data processing steps are transparent and reproducible.

## **Why This Matters:**

Daily-level data is especially prone to noise, glitches, and occasional outages from exchanges or APIs. By building robust quality checks into our pipeline, we minimize the risk of basing our analysis on flawed data. This diligence is crucial for producing trustworthy visualizations, accurate forecasts, and meaningful insights.

**Reflection:**

We have learned that data quality is not a one-time task but an ongoing process. Regular validation, clear documentation, and a healthy skepticism toward “too good to be true” numbers have helped us build a dataset we can trust and that others can build on in the future.

# **Data Analysis and Visualization**

With a robust, daily dataset for both Bitcoin and Ethereum, our project moves from data engineering to the exciting world of analysis and visualization. This is where we transform raw numbers into actionable insights and compelling stories about the crypto market’s behavior.

## **Exploratory Data Analysis (EDA):**

We began by exploring the cleaned datasets (btc\_cleaned.csv and eth\_cleaned.csv) using Python’s pandas. Our EDA included:

* **Time Series Plots:** Visualizing price movements at the daily level revealed micro-trends, flash crashes, and periods of intense trading activity that are invisible in daily data.
* **Volume Patterns:** Analyzing trading volume at a granular level helped us spot liquidity surges, market openings, and the impact of global events on trading behavior.

## **Comparative Analysis: Bitcoin vs. Ethereum:**

* **Volume & Momentum:** Line and Stacked Column chart of both BTC and ETH helped us to compare daily trading volume vs 30 SMA price will show us the price variation backed by healthy liquidity.
* **Lag Analysis:** By shifting one asset’s time series against the other, we explored whether price changes in Bitcoin tend to lead or lag those in Ethereum a key insight for traders and researchers.
* **Year over Year Return Comparison:** Clustered Column chart helped us compare both BTC and ETH closes with year that and helped us check which coin outperformed in each year.

## **Advanced Visualizations:**

* **Heatmaps:** We created heatmaps to visualize trading activity by daily revealing when the market is most active and when significant price changes are most likely to occur.

## **Forecasting and Modeling:**

* **Short-Term Forecasts:** Using ARIMA and Prophet models, we generated short-term price forecasts based on recent daily-level trends. These models helped us understand the predictability (or unpredictability) of crypto prices at high frequency.
* **Anomaly Detection:** We applied statistical techniques to flag unusual price movements or volume spikes, which could indicate market manipulation, technical glitches, or breaking news.

## **Key Insights:**

* Daily level data reveals a much more dynamic and “alive” market than daily data ever could.
* Bitcoin and Ethereum are highly correlated, but their relationship can shift rapidly, especially during periods of market stress or innovation.
* High-frequency analysis uncovers patterns such as recurring volatility at certain times of day that can inform trading strategies and risk management.

**Reflection:**

This phase of the project was both challenging and rewarding. The sheer volume of data required efficient coding and visualization techniques, but the payoff was a much deeper understanding of how the crypto market really works, day by day.

## **Prices of Bitcoin & Ethereum over the year**

A graph showing the price of a stock market

AI-generated content may be incorrect.

## **Return Percentage Over the Years by Horizon Label BTC/ETH and Coin**

A graph showing the number of percentages

AI-generated content may be incorrect.

## **Average Daily Trading Volume of Bitcoin vs 30 Day Moving Average**

A graph with a line going up

AI-generated content may be incorrect.

## **Average Daily Trading Volume of Ethereum vs 30 Day Moving Average**

A graph with a line going up

AI-generated content may be incorrect.

**Monthly Trading Volume Heatmap**

**A table with numbers and numbers

AI-generated content may be incorrect.**

# **Extension**

As our project evolved, it became clear that the tools, pipelines, and insights we developed could be scaled and adapted in many powerful ways. The high-frequency, daily level approach we used for Bitcoin and Ethereum is just the beginning there is a vast landscape of possibilities for future work and real-world application.

## **Expanding to More Assets:**

Our current pipeline is designed to be modular and flexible. With minor adjustments, it can ingest and process data for additional cryptocurrencies. This would allow for broader market analysis, sector comparisons, and the study of correlations across the entire crypto ecosystem.

## **Increasing Data Granularity and Real-Time Analytics:**

While daily data is already highly granular, the crypto market operates in real time. By adapting our scripts to handle second-level we could capture even more detailed market microstructure. This would be especially valuable for algorithmic trading research, market making, or studying the impact of high-frequency trading bots.

## **Scaling Data Storage and Processing:**

As the volume of data grows, so do the demands on storage and computation. We are exploring:

* **Time-Series Databases:** Using Relational Database to efficiently store and query massive time-series datasets.

## **Real-Time Dashboards and Alerts:**

A natural next step is to build interactive dashboards using tools like PowerBI. These dashboards could display heatmaps, for unusual market activity making our insights accessible to traders, researchers, and the broader crypto community.

**Reflection:**

The skills and systems we’ve built are a solid foundation for ongoing research and practical applications in the fast-moving world of crypto. Whether for academic study, trading, or market monitoring, our project is ready to grow and adapt to new challenges.

# **Allocation Project Team Roles**

A project of this scale and complexity succeeds only through effective teamwork and clear division of responsibilities. Our team leveraged each member’s strengths, ensuring that every aspect of the project from data engineering to analysis and reporting was handled with care and expertise.

## **Team Member Roles and Contributions:**

* **Krishna Vamsi Cherukupalli:**

Krishna led the data engineering efforts, focusing on the development and automation of the data extraction and cleaning pipelines. He was responsible for writing and maintaining the core Python scripts that fetch, validate, and standardize high-frequency trading data from multiple sources. Krishna also contributed to the design of the ETL (Extract, Transform, Load) workflow and ensured that our datasets were always up-to-date and analysis ready.

* **Lohith Reddy Danda:**

Lohith specialized in data analysis and visualization and, quality assurance. He explored the cleaned datasets, performed statistical analyses, and created a wide range of visualizations from time series plots to heatmaps and candlestick charts. Lohith also took the lead in drafting the project report, ensuring that our findings were clearly communicated and well-supported by data.

* **Paras Rupani:**

Paras managed data storage and project coordination. He oversaw the organization and backup of data files, implemented version control practices using GitHub.

## **Team Communication and Workflow:**

* We held regular meetings (both virtual and in-person) to discuss progress, troubleshoot issues, and plan next steps.
* Responsibilities were clearly defined, but we also supported each other and collaborated on overlapping tasks, especially during critical phases like data cleaning and final report preparation.

## **What Worked Well:**

* Clear division of labor and open communication helped us stay organized and efficient.
* Regular check-ins allowed us to quickly identify and resolve challenges.

## **What We would Improve Next Time:**

* Implementing more automated reminders and progress tracking could further streamline our workflow.
* Allocating extra time for unexpected data issues or technical bugs would help reduce last-minute stress.

**Reflection:**

Our teamwork was a key factor in the project’s success. By leveraging each member’s strengths and maintaining a supportive, communicative environment, we were able to tackle complex challenges and deliver a high-quality result.

# **Project Timeline**

A well-planned timeline is essential for managing a complex, data-driven project especially when working with high-frequency data and multiple team members. Our timeline provided structure, accountability, and flexibility, helping us stay on track from initial planning to final submission.

## **Project Timeline Overview:**

| **Date** | **Deliverable** | **Responsible** |
| --- | --- | --- |
| July 10 | Data sources identified | Krishna Vamsi Cherukupalli |
| July 12 | Data extraction scripts developed | Krishna Vamsi Cherukupalli |
| July 15 | Initial data collection completed | Krishna Vamsi Cherukupalli |
| July 17 | Data cleaning and validation | Lohith & Krishna |
| July 20 | Data integration and storage setup | Paras Rupani |
| July 22 | Exploratory data analysis (EDA) | Lohith Reddy Danda |
| July 24 | Advanced analysis & visualization | Lohith Reddy Danda |
| July 26 | Draft report prepared | All |
| July 29 | Final review and edits | All |
| July 31 | Report submission | All |

## **How We Managed Our Timeline:**

* We used a shared Google Sheet and GitHub Issues to map out deliverables, assign responsibilities, and track progress.
* Weekly meetings served as checkpoints to review our status, adjust plans, and ensure we were meeting our milestones.
* Buffer time was built in for unexpected challenges, such as data quality issues or technical bugs.

## **Lessons Learned:**

* Early planning and clear task assignments helped us avoid last-minute rushes.
* Regular communication and progress tracking were crucial for staying aligned, especially when working remotely.
* Flexibility was important when we encountered unexpected data issues, we were able to adjust our schedule and reassign tasks as needed.

**Reflection:**

Our timeline was more than just a schedule; it was a living document that guided our work, kept us accountable, and helped us celebrate our progress. By planning ahead and staying organized, we were able to deliver a comprehensive, high-quality project on time.

# **References**

* *yfinance. (2025, July 6).* PyPI. <https://pypi.org/project/yfinance/>
* *pandas’ documentation — pandas 2.3.1 documentation*. (n.d.). <https://pandas.pydata.org/docs/>
* *NumPY Documentation.* (n.d.). <https://numpy.org/doc/>
* *David-Engel. (n.d.). Microsoft ODBC Driver for SQL Server - ODBC Driver for SQL Server*. Microsoft Learn. <https://learn.microsoft.com/sql/connect/odbc/>
* *Prophet. (n.d.). Prophet.* <https://facebook.github.io/prophet/>
* *prophet. (2025, May 30).* PyPI. <https://pypi.org/project/prophet/>
* *MiguelMyersMS. (n.d.). Combo chart in Power BI - Power BI. Microsoft* Learn. <https://learn.microsoft.com/power-bi/visuals/power-bi-visualization-combo-chart>
* *Time Series analysis tsa - statsmodels 0*.14.4. (n.d.). <https://www.statsmodels.org/stable/tsa.html#exponential-smoothing>
* *pandas.DataFrame.ewm — pandas 2.3.1 documentation.* (n.d.). <https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.ewm.html>
* *pandas.Series.pct\_change* — pandas 2.3.1 documentation. (n.d.). <https://pandas.pydata.org/docs/reference/api/pandas.Series.pct_change.html>
* *WilliamDAssafMSFT. (n.d.). TRUNCATE TABLE (Transact-SQL) - SQL Server.* Microsoft Learn. <https://learn.microsoft.com/sql/t-sql/statements/truncate-table-transact-sql>