CS 418 Project Presentation

Ethereum Blockchain Transaction Analysis

Team - 2

Instructor - Dr. Sathya Ravi

Team Members

Nastaran Darabi - Github

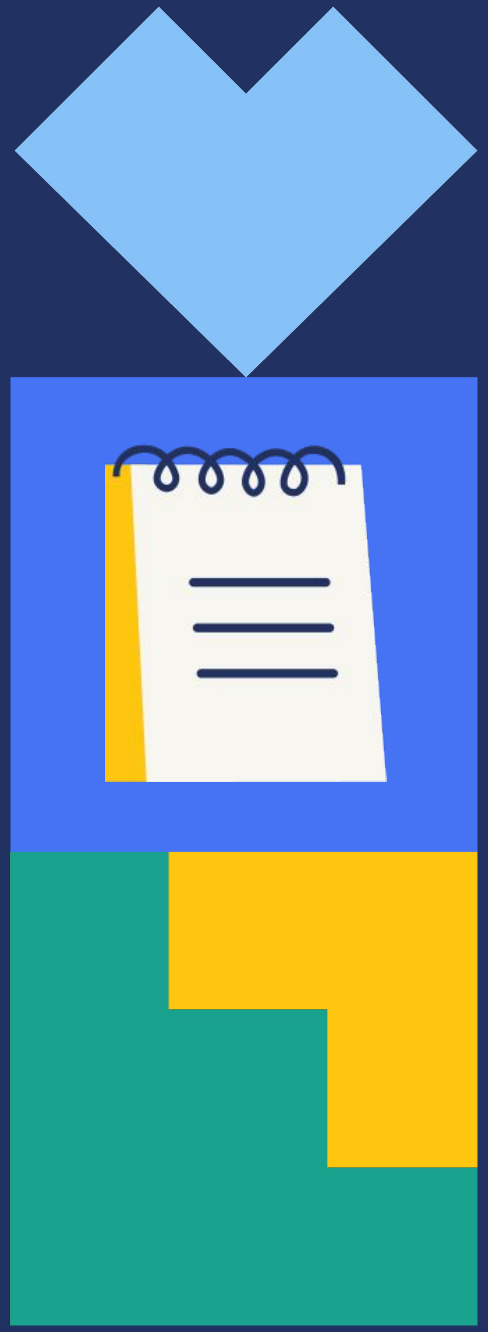
Shreyash Kadam - Github

Sanmitha Shetty - Github

Dhiraj Shelke - Github

Frank Wang - Github

Victor Escudero - Github



Project Introduction

The Ethereum blockchain is a decentralized platform enabling transactions and smart contracts.

This project investigates Ethereum transactions to derive insights into patterns, trends, and financial implications.

Focus areas:

- Transaction analysis by type and fee
- Prediction metrics for Ethereum price movements
- Exploration of blockchain dynamics through data analysis

Problem Statement

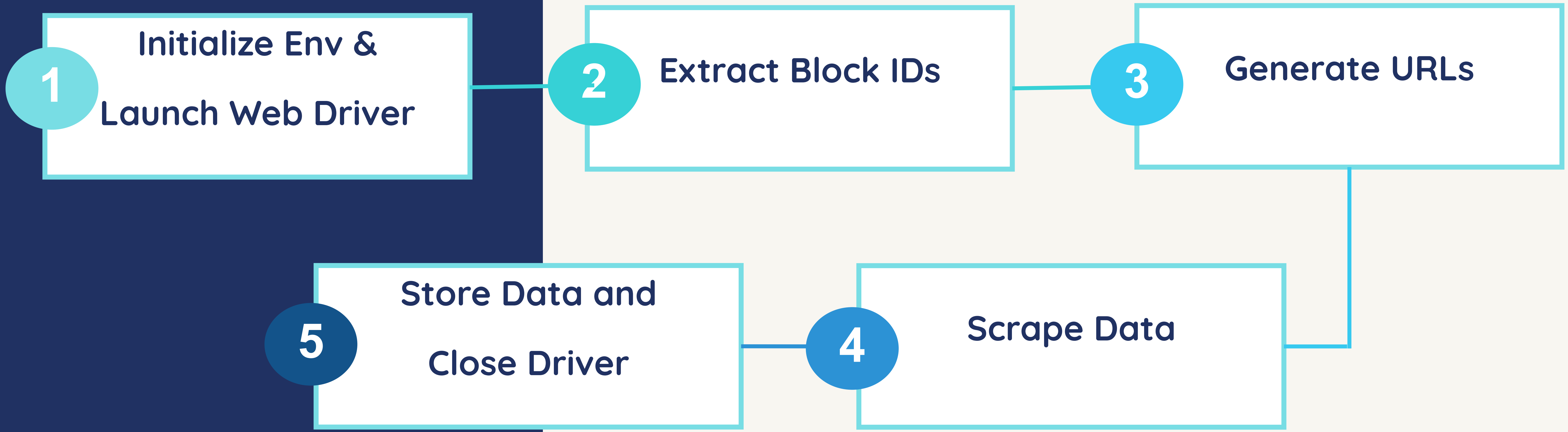
- Blockchain transactions hold significant value for financial systems.
- Ethereum, being a major blockchain network, facilitates numerous transactions daily.
- Key Questions:
  a. What patterns or trends can we observe in Ethereum transactions?
  b. Can transaction metrics predict Ethereum price movements?
- Motivation: Understanding these trends helps blockchain users, developers, and financial analysts.

Methodology

1. Data Collection and Web Scraping
2. Data Cleaning and Preprocessing
3. Data Visualization & Distribution Analysis
4. Exploratory Data Analysis
5. Machine Learning Analysis

Data Collection

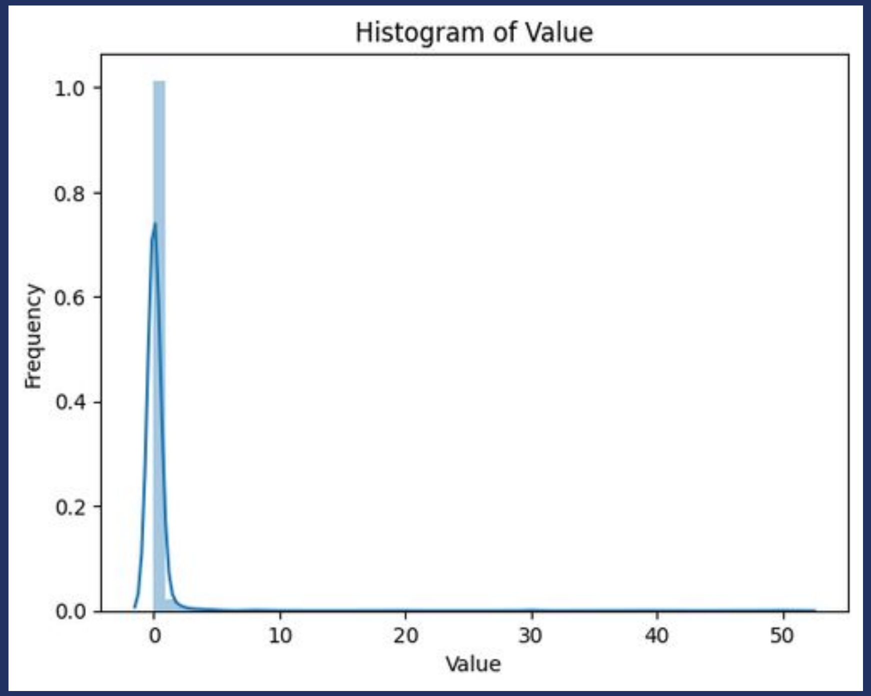
- Data Source: The data is collected from Etherscan.io (https://etherscan.io/txs) using Selenium,
- Initial Observations: The collected data enables an analysis of transaction patterns, including frequency, transaction values, and fees.
- Using Selenium: we automate the collection of transaction data.

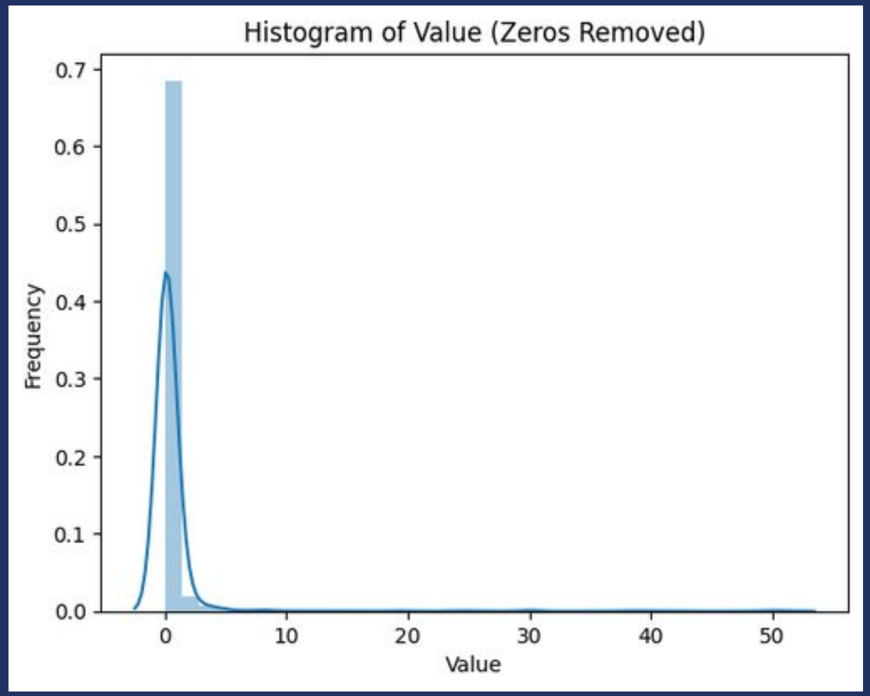


Data Cleaning

Preprocessing Steps:

- Removed invalid/missing records.
- Filtered for recent transactions.
- Extracted and converted relevant numeric values.





txnHash	method	block	age	from	to	value	txnFee
0x036416:	Transfer	21124918	24 secs ago	javascript:;	0x13F2241	0.09267941ε	0.00012626
0x7afd04c	Commit Blob	21124918	24 secs ago	0x5050f69a	0xFf00000(	0 ETH	0.00023959
0xbacf5a(	Propose Block	21124918	24 secs ago	0x9084ee7	0xeCEc54:	0 ETH	0.00137946
0xabe173:	Transfer	21124918	24 secs ago	0xa62a9ed	0x948BFF!	1 ETH	0.00015776

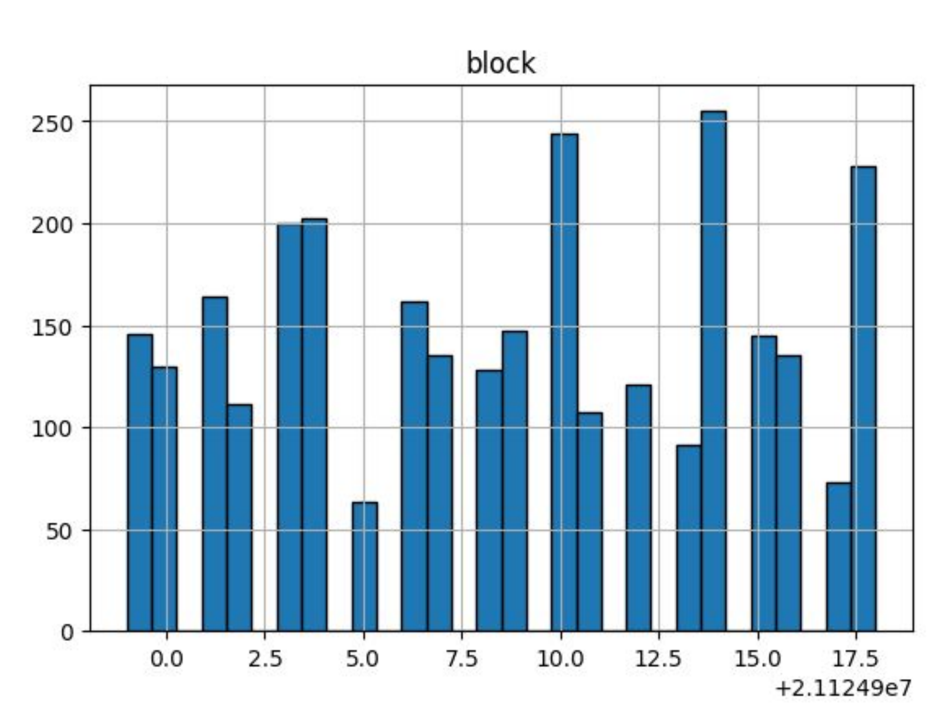
Data Visualization and Distribution Analysis

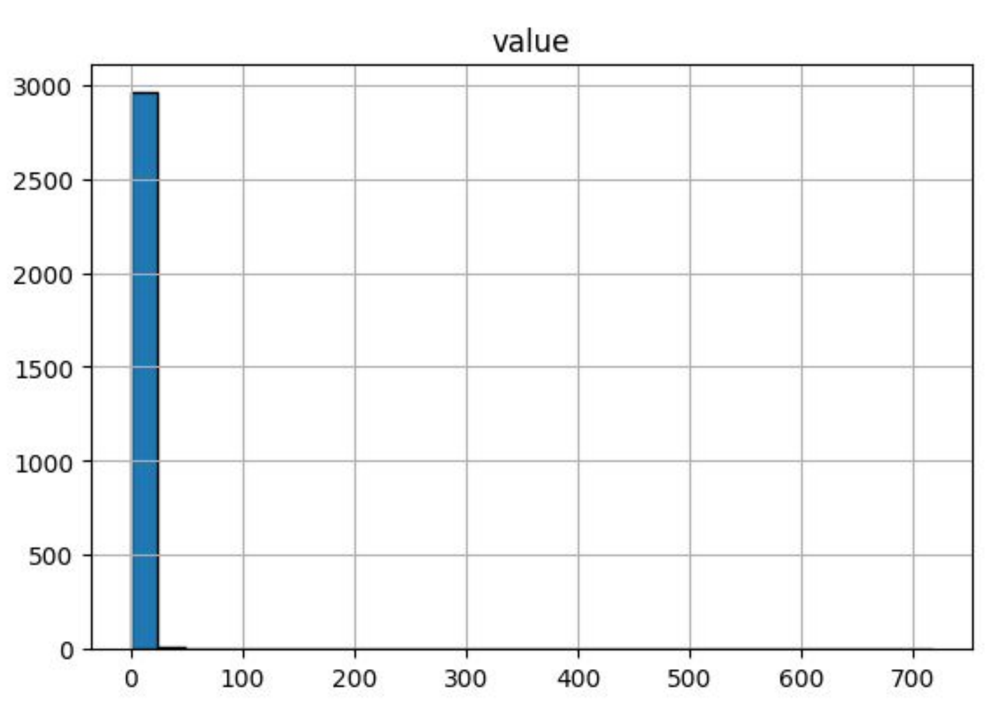
Histogram Bin Sizing: Used square root of dataset size to determine optimal bin size

Histograms of Transaction Values and Fees:

Transaction Values: Analyzed spread and frequency to identify dominant transaction sizes.

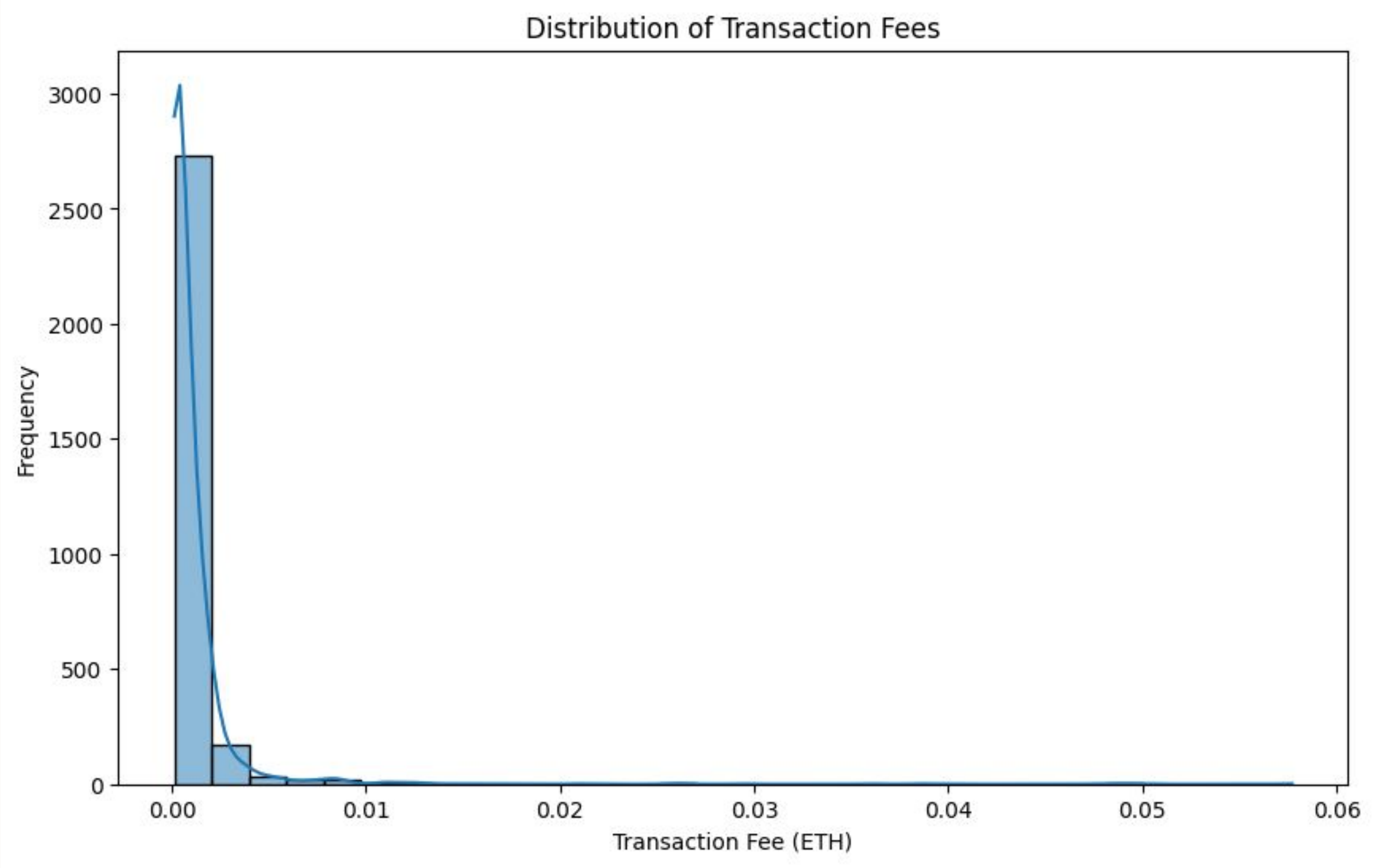
Transaction Fees: Evaluated fee distribution to uncover trends and patterns.





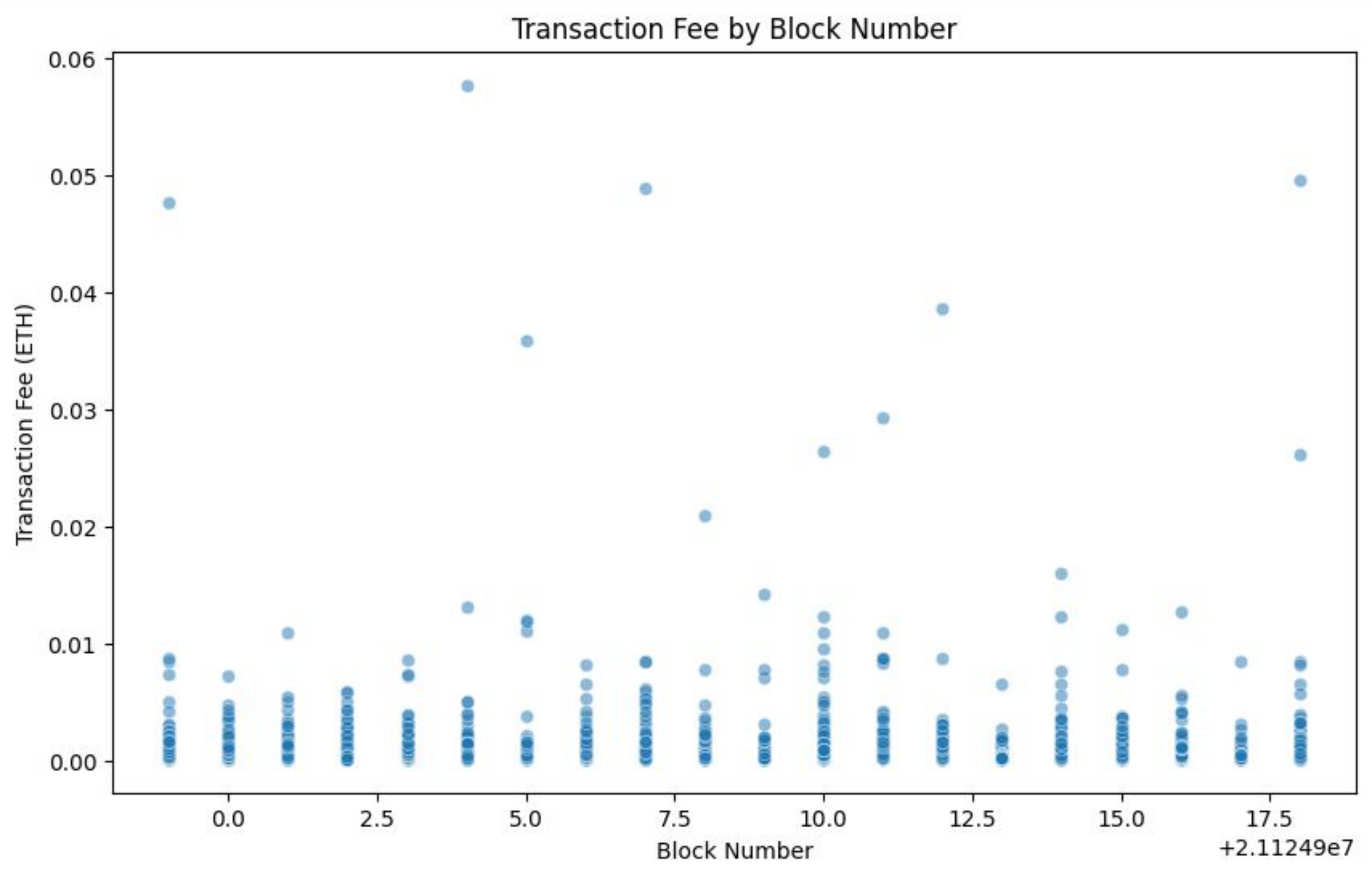
Exploratory Data Analysis

- **Transaction Fee Distribution: A histogram with KDE (kernel density estimate) visualizes the spread of transaction fees, highlighting typical values and outliers.**



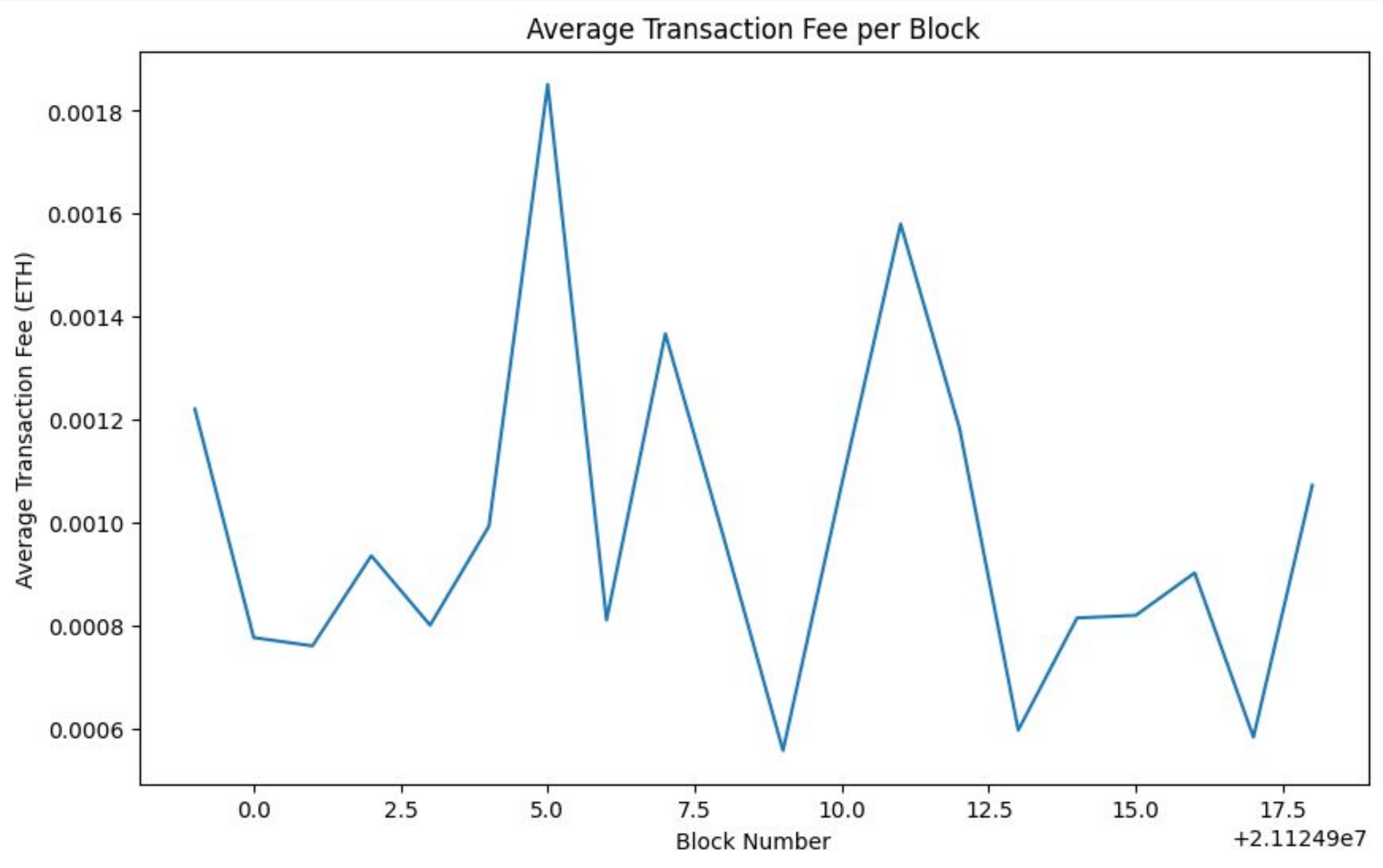
Exploratory Data Analysis

- **Block Number Correlation:** A scatter plot explores the relationship between transaction fees and block numbers, allowing us to observe if fees vary significantly by block.



Exploratory Data Analysis

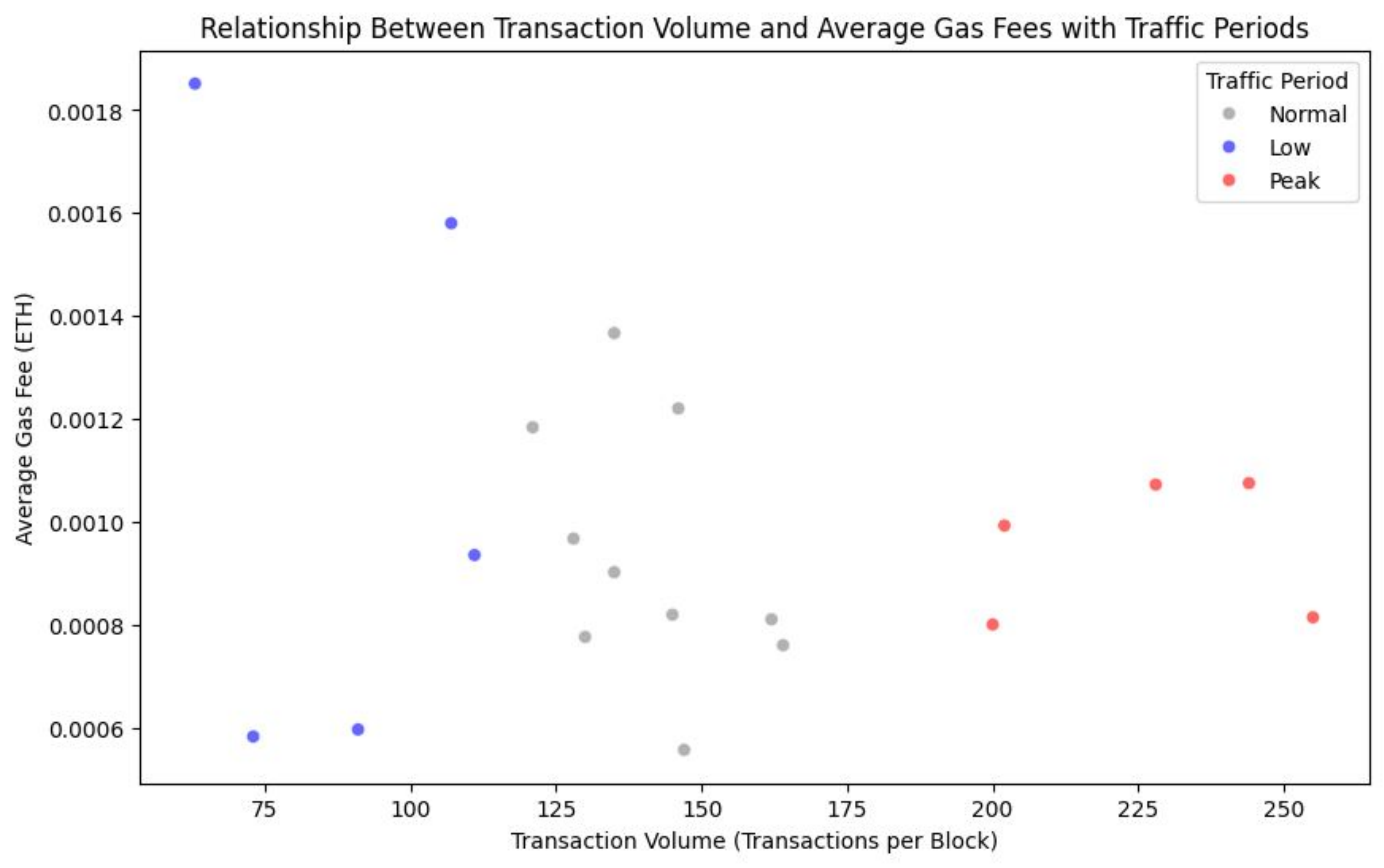
- **Average Transaction Fee per Block:** A line plot shows the trend of average transaction fees over blocks, which can help identify any temporal trends in transaction costs.



Hypothesis Testing Visualization

Hypothesis:

"Higher gas fees indicate higher transaction volume and network congestion."



Hypothesis Testing Visualization

Hypothesis: “Higher gas fees indicate higher transaction volume and network congestion.”

- Findings: The plot revealed that while "Peak" transaction volumes often align with higher gas fees, there are exceptions. Some "Low" transaction volume periods also show high fees, and "Normal" periods exhibit a wide range of gas fees with no consistent pattern.

- Conclusion: Although transaction volume influences gas fees, additional factors like network congestion, block demand, or fee structure changes likely contribute as well. Further statistical analysis or more data is needed to explore these relationships.

Machine Learning Analysis

- Feature Engineering: New features such as rolling averages for transaction volumes and fees, and transaction value ratios to highlight spikes and anomalies.

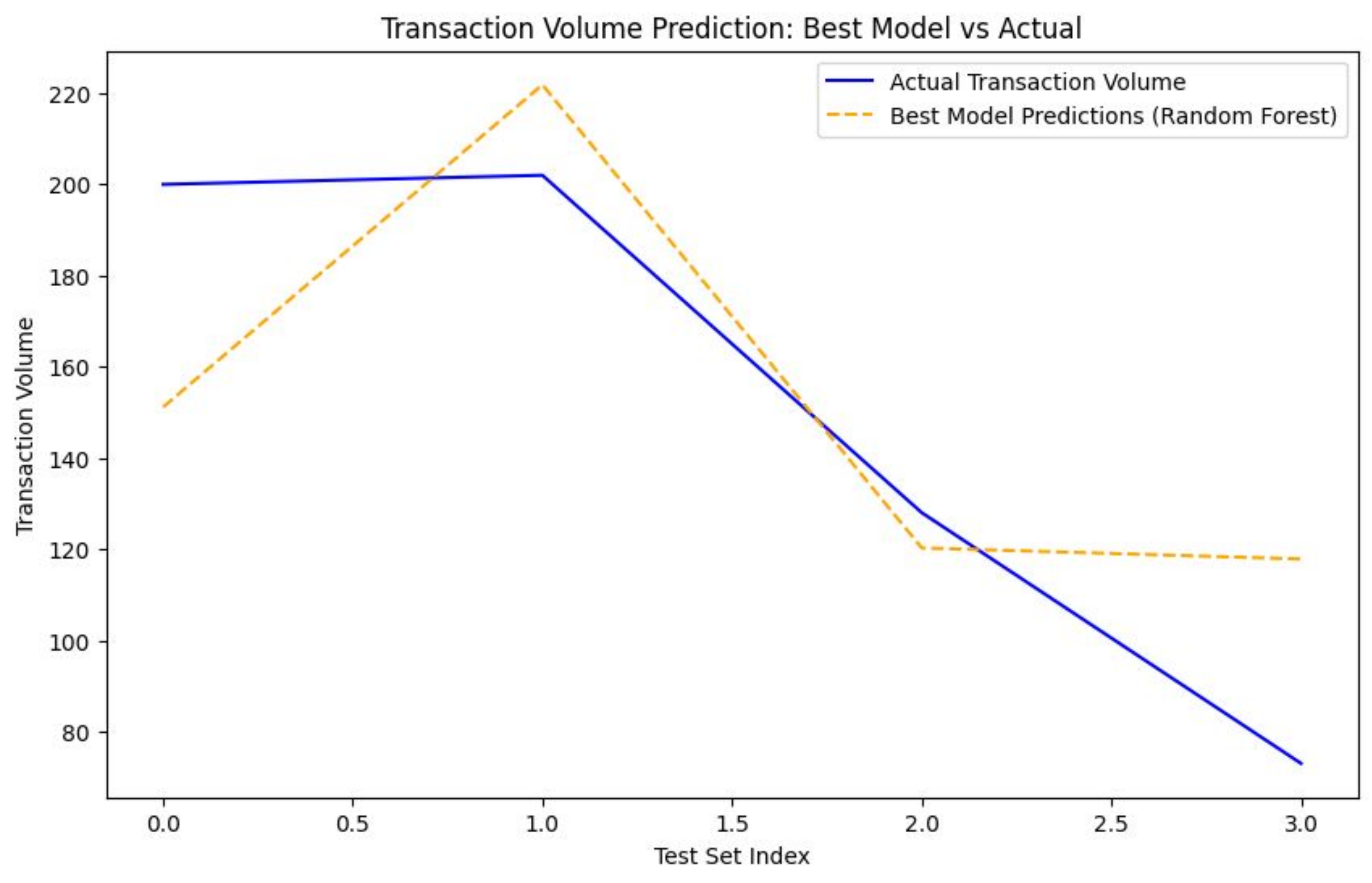
- Explored Models:
  - Random Forest Regressor: Captures nonlinear relationships.
  - Gradient Boosting Regressor: Effective for tabular data.
  - Lasso and Ridge Regression: Prevent overfitting on smaller datasets.
  - Hyperparameter Tuning: Used cross-validation to optimize model performance.

- Evaluation Metrics: Evaluated models using MAE, MSE, RMSE, and R^2 for comprehensive accuracy assessment.

Results

Key Observation:

- Actual transaction volume shows a clear upward trend.
- Random Forest predictions remain flat and significantly lower.



The plot reveals a noticeable gap between Actual Transaction Volume (blue line) and Model Predictions (orange dashed line).

Takeaways

Data Insights:

- Skewed distributions

Machine Learning Observations:

- Feature engineering enhanced data but was insufficient to address predictive discrepancies.
- Random Forest and Gradient Boosting models showed potential but struggled to capture the trends accurately.

Challenges

Data Quality:

- High prevalence of outliers and zero-value
- Imbalanced data

Model Limitations:

- Difficulty in capturing non-linear
- Computational complexity

Feature Engineering:

- Additional relevant features that improve predictive accuracy.

Thank You!