PROGRESS REPORT

Project Title: Analysis of Trends in the NFT Market: Valuable NFTs, Price Factors, and Wash Trading

Group Members: Aslı AYAZ, Zehra Bengisu DOĞAN, Zeynep ÇALAPKULU

Submission Date: 29/04/2025

1. Introduction

In recent years, the NFT (Non-Fungible Token) market has emerged as a significant component of the digital economy, experiencing rapid growth and attracting substantial attention from investors, artists, and technologists. NFTs represent unique digital assets verified using blockchain technology, enabling the ownership and exchange of digital art, collectibles, and other virtual items. The surge in NFT popularity has led to a dynamic and often volatile market environment, characterized by significant price fluctuations and trading volumes.

However, this burgeoning market is not without challenges. One of the primary concerns is the prevalence of manipulative practices such as wash trading, where traders buy and sell the same asset to create a misleading appearance of high demand and inflated prices [1]. Studies have indicated that a substantial portion of NFT trading volume may be attributed to such activities, raising questions about market integrity and the true value of these digital assets [2].

To analyze these phenomena, we utilize the "OpenSea NFT Sales 2019–2021" dataset [3], which provides comprehensive transaction data from one of the leading NFT marketplaces. This dataset includes detailed information on sales transactions, asset metadata, pricing, and participant addresses, offering a rich resource for examining market trends, price determinants, and potential indicators of wash trading.

Our project aims to dissect the NFT market's complexities by identifying key factors influencing NFT valuations, detecting patterns indicative of wash trading, and understanding the temporal dynamics of NFT sales. Through statistical analysis, clustering, and time-series modeling, we seek to uncover insights into the behaviors driving this digital marketplace.

By shedding light on these aspects, our research endeavors to contribute to a more transparent and informed understanding of the NFT ecosystem, benefiting stakeholders ranging from individual collectors to institutional investors

2. Determining Algorithms

In this phase of the study, we have identified a comprehensive set of algorithms and analytical methods that are best suited for exploring the dynamics of the NFT market. These algorithms have been selected based on their ability to capture temporal patterns, extract meaningful features, detect irregularities, and understand underlying relationships between data variables. The methods are designed to address our core research questions and provide a multidimensional understanding of how NFT prices behave, what factors influence them, and how anomalies such as wash trading can be detected. Leveraging techniques from time series analysis, unsupervised learning, statistical modeling, and graph theory, this framework aims to offer both depth and flexibility in analyzing the NFT ecosystem.

2.1 Time Series Analysis

To explore how NFT sales and prices fluctuate over time and to uncover underlying temporal patterns, we will employ time series analysis techniques. These methods are essential for identifying trends, seasonal effects, and market shifts, providing insight into the long-term behavior of the NFT market [4].

- **Moving Average:** This is a fundamental smoothing technique that helps reduce noise in time series data by averaging data points over a specified period. It assists in identifying long-term movements and suppressing short-term fluctuations.
- **Trend Analysis:** Trend analysis involves determining the general direction in which data points are moving over time. We will visualize NFT sales and price changes using line graphs and time series heatmaps to detect persistent upward or downward trajectories that might reflect market interest or hype cycles.

2.2 Identifying Factors Affecting NFT Prices

Understanding which features influence NFT pricing is critical to market analysis. We will use a combination of feature selection, correlation analysis, and regression modeling to explore these relationships.

- **Feature Selection and Correlation Analysis:** By computing correlations among attributes such as asset collection names, number of sales, token standards, and transaction metadata, we aim to determine which factors most strongly impact NFT valuation.
- **Pearson Correlation Coefficient:** This statistical measure evaluates the linear correlation between two variables, ranging from -1 to +1. It will be used to assess the strength and direction of relationships between NFT features and price outcomes [5].
- **Regression Models:** Linear and multiple linear regression models will be used to predict NFT prices based on selected variables. These models can quantify the degree to which each feature contributes to price variation, helping to explain how multiple independent variables simultaneously affect an outcome variable.

2.3 Examining the Most Valuable NFTs in Collections

To identify the most valuable NFTs and understand the attributes they share, we will use unsupervised machine learning algorithms, specifically clustering and frequent pattern mining techniques.

- **K-Means Clustering:** This algorithm partitions the dataset into 'k' clusters based on feature similarity, enabling the identification of NFTs that belong to the same group in terms of attributes like price, category, or sales frequency [6]. The grouping can reveal which traits are common among high-value NFTs.
- **Frequent Pattern Mining (FP-Growth):** FP-Growth is an efficient algorithm for mining frequent itemsets without candidate generation [7]. In our context, it will be used to identify commonly co-occurring features (e.g., file type, creator, metadata tags) among top-selling NFTs, helping us understand what makes them valuable.

2.4 Detecting Wash Trading

Wash trading is a manipulative practice where traders buy and sell the same NFT multiple times to artificially inflate its value and perceived popularity. Detecting such anomalies is crucial for ensuring market transparency.

- **Graph Analysis:** Graph-based methods allow us to model buyer-seller relationships and detect suspicious patterns such as repeated transactions between the same addresses. If time permits, social network analysis metrics like centrality and clustering coefficients will be explored to identify potential fraud rings [8].
- **Outlier Detection:**
  - **Isolation Forest:** This algorithm isolates anomalies rather than profiling normal data. Since outliers are easier to isolate, fewer steps are required to separate them. It is efficient for high-dimensional datasets and performs well in identifying unusual trading behaviors [9].
  - **Local Outlier Factor (LOF):** LOF measures the local deviation of a given data point with respect to its neighbors. It is useful for identifying contextual anomalies where trading behavior differs significantly from similar transactions.

2.5 Analyzing Seasonal and Periodic Changes in the NFT Market

Seasonal patterns in digital asset markets can result from investor behavior, marketing events, or global economic cycles. Detecting such trends allows for a deeper understanding of market psychology and timing.

- **Seasonal Decomposition (STL):** STL decomposes time series into seasonal, trend, and residual components using Loess smoothing, providing a clearer view of cyclic behavior [10].

- **Time Series Heatmaps:** These visualizations will depict the frequency or volume of NFT transactions across various time intervals (e.g., monthly, quarterly), aiding in the identification of temporal spikes or drops.
- **Clustering per Time Period:** By segmenting the dataset by time (e.g., quarters, years) and applying clustering, we can observe how NFT characteristics and market behavior evolve over different periods.

2.6 Examining NFT Price Volatility

Volatility measures how much and how quickly prices change over time, and is a key indicator of risk and investor sentiment.

- **Volatility Calculation:** We will calculate the standard deviation and the coefficient of variation (CV) of NFT prices to assess overall market volatility. CV, which is the ratio of the standard deviation to the mean, is especially useful for comparing relative variability across different asset classes.
- **Rolling Window Analysis:** This method applies statistical metrics over a sliding window of time, allowing us to observe how volatility and other features evolve. It is particularly useful for monitoring the dynamic nature of markets.
- **Visualization:** Line charts and volatility bands will be used to present volatility trends over time, helping to highlight periods of instability or rapid market shifts.

3. Data Preprocessing and Cleaning

3.1 Handling Missing Data and Standardization

In this phase, we addressed missing values and standardized key attributes to ensure the dataset is consistent, complete, and ready for reliable analysis. Each column was handled carefully to avoid introducing bias or distorting the original data distribution.

- **asset.name**: If this field was missing but the asset.collection.name was available, we created a placeholder by combining the collection name with a generic term (e.g., “Unnamed from X Collection”). If both values were missing, we assigned a default value such as "Unnamed Asset" to preserve dataset completeness.
- **asset.collection.name**: For entries lacking a collection name, we attempted to infer it using the asset.name (if available). In cases where neither field was present, a generic placeholder like "Unknown Collection" was used.
- **payment_token.name**: Missing values in this field were filled with the most frequently occurring token in the dataset. This approach ensures minimal deviation from the overall distribution and avoids introducing rare or anomalous tokens into missing fields.
- **payment_token.usd_price**: If the USD price of the token was missing, we calculated it using the total_price and the known exchange rate of the token (based on pre-defined values, such as Ether-to-USD). This conversion ensured consistency in monetary values across the dataset.
- **seller.user.username**: When usernames were missing, we matched the associated seller.address with any known username from other records. If no match was found, we

generated a unique placeholder username based on the address, maintaining the ability to track sellers even without explicit usernames.

- **Categorical Data Standardization**: We standardized text values in columns such as asset.name, asset.collection.name, and payment_token.name to resolve inconsistencies due to typographical variations (e.g., "ETH" vs. "Eth"). This ensured uniform categorization during analysis.

These steps were crucial in transforming the raw dataset into a clean and reliable format suitable for further exploration and modeling.

3.2 Data Cleaning

We also eliminated certain attributes that were irrelevant or unhelpful for our analysis:

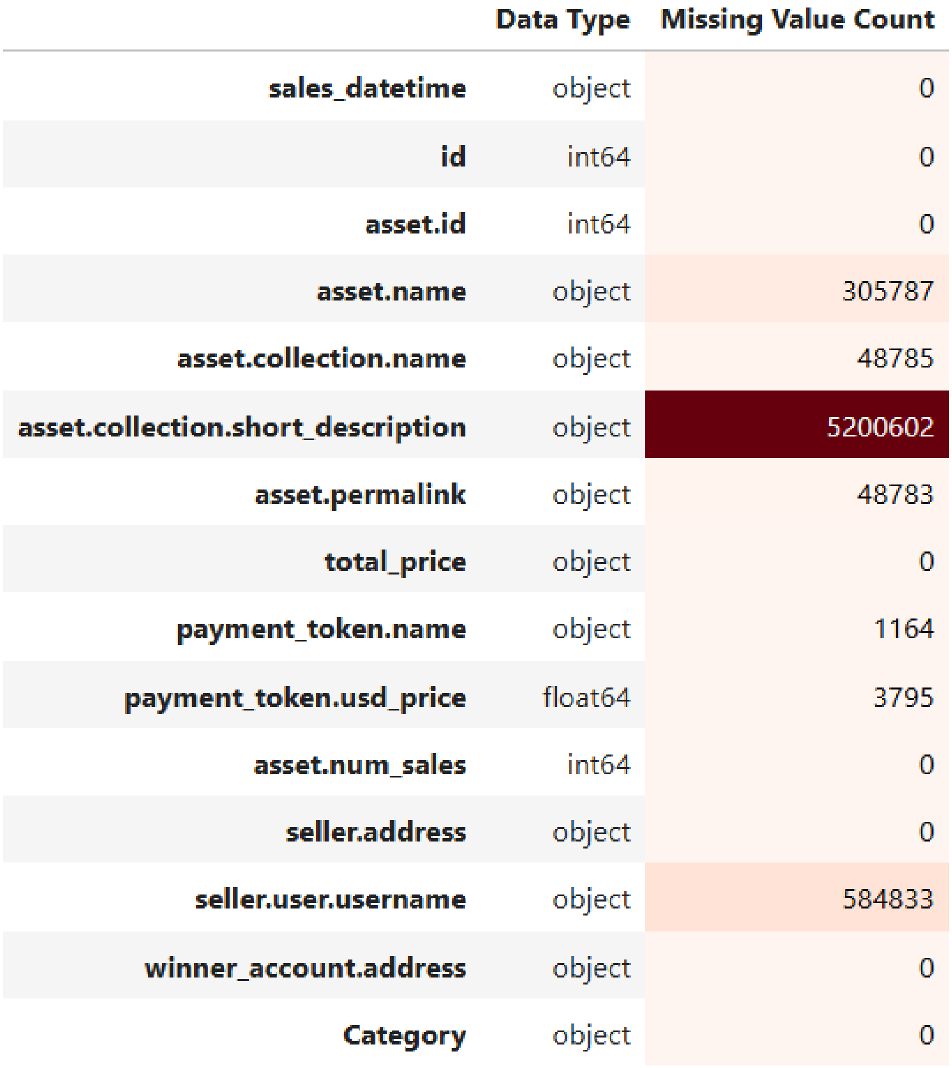
- **asset.collection.short_description**: This field contained a high proportion of missing or meaningless values. Since it lacked analytical value and cluttered the dataset, it was removed entirely.
- **asset.permalink**: As this column only contained hyperlinks to asset listings, and offered no analytical insight, it was excluded from the dataset to streamline preprocessing and focus on more relevant variables.

4. Initial Analysis

In this phase, we aimed to explore the structure, completeness, and underlying patterns in the dataset, which consists of 5,252,255 records and 15 attributes. Our goal was to identify potential insights and better understand how the data is distributed across different features. We carried out several key steps as part of this initial exploratory analysis.

Missing Values and Cleaned Attributes: We checked the dataset for missing values (nulls), as well as which attributes had been filled or removed during the data preprocessing phase. This provided insight into the completeness of the data and the steps we took to clean it.

Identifying Numerical and Categorical Columns: We categorized the columns into numerical and categorical types. This distinction allowed us to perform appropriate statistical and frequency analysis based on the data types.



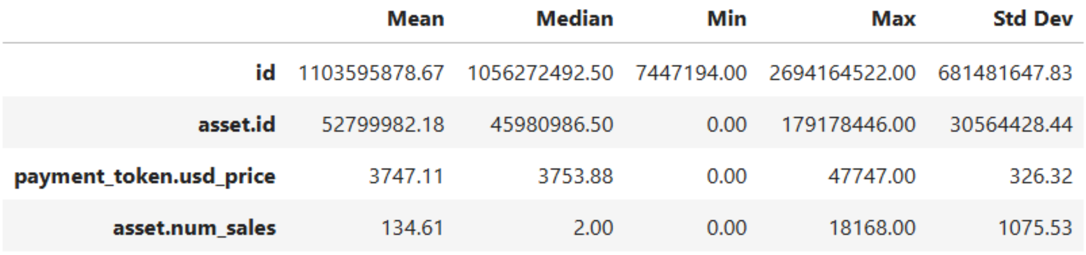
	Data Type	Missing Value Count
sales_datetime	object	0
id	int64	0
asset.id	int64	0
asset.name	object	305787
asset.collection.name	object	48785
asset.collection.short_description	object	5200602
asset.permalink	object	48783
total_price	object	0
payment_token.name	object	1164
payment_token.usd_price	float64	3795
asset.num_sales	int64	0
seller.address	object	0
seller.user.username	object	584833
winner_account.address	object	0
Category	object	0

Fig. 1. Table displaying data types and missing value counts for each attribute, with columns containing high numbers of missing values highlighted in darker red.

From the table above, several observations can be made:

- The column asset.collection.short_description has missing values in over 5.2 million rows, indicating it was largely unused in the dataset. Due to its lack of contribution to meaningful analysis, this attribute was removed during the data cleaning phase.
- asset.name, asset.collection.name, seller.user.username, and payment_token related fields also contain missing values, which were addressed through logical imputations during the preprocessing stage (as discussed in Section 3.1).
- Most attributes are of type object, indicating they are either string-based categorical values or identifiers, while a few are numeric (int64, float64), enabling statistical analysis.

Basic Statistical Analysis for Numerical Columns: For the numerical columns, we computed basic statistics such as the mean, median, minimum, maximum, and standard deviation. These metrics provided a better understanding of the distribution and spread of numerical data in the dataset.

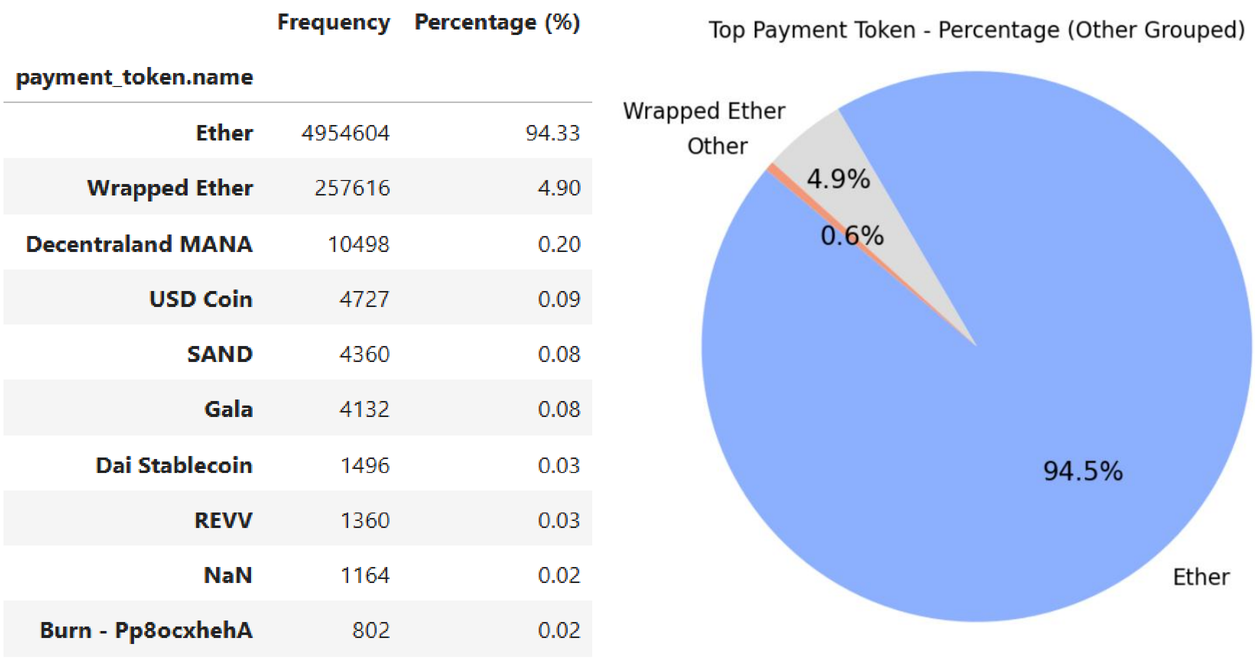


	Mean	Median	Min	Max	Std Dev
id	1103595878.67	1056272492.50	7447194.00	2694164522.00	681481647.83
asset.id	52799982.18	45980986.50	0.00	179178446.00	30564428.44
payment_token.usd_price	3747.11	3753.88	0.00	47747.00	326.32
asset.num_sales	134.61	2.00	0.00	18168.00	1075.53

Fig. 2. Basic Statistical Analysis for Numerical Columns.

- The id and asset.id columns are skewed right, as evident from the mean being significantly greater than the median. This reflects a long tail of high-value IDs, likely representing newer entries or outliers.
- The payment_token.usd_price has a relatively low standard deviation compared to its mean, suggesting that most token prices are clustered around the average.
- A striking point is the distribution of asset.num_sales, with a median of 2 but a mean of 134.61, indicating that while most NFTs were sold very few times, a small group of highly traded assets skewed the distribution. This long-tail behavior is typical in NFT marketplaces.

Frequency of Categorical Variables: We analyzed the frequency of categorical variables, such as identifying which tokens were used and which ones were predominant in the dataset. This helped us understand the distribution of token usage across the NFT sales.



	Frequency	Percentage (%)
payment_token.name		
Ether	4954604	94.33
Wrapped Ether	257616	4.90
Decentraland MANA	10498	0.20
USD Coin	4727	0.09
SAND	4360	0.08
Gala	4132	0.08
Dai Stablecoin	1496	0.03
REVV	1360	0.03
NaN	1164	0.02
Burn - Pp8ocxhehA	802	0.02

Fig. 3. Table and graph showing the frequency of categorical variables.

- Ether is the overwhelmingly dominant token, used in more than 94% of transactions.
- The next most used, Wrapped Ether, accounts for nearly 5%, while other tokens like Decentraland MANA or USD Coin are marginal.
- A total of 198 unique tokens were observed, but the majority have negligible usage rates (less than 0.02%), indicating a highly concentrated market dominated by a few currencies.

Unique Collections and NFTs: We examined how many unique collections and NFTs were present in the dataset. This analysis helped us understand the diversity of the NFT market represented in the dataset.

Number of unique collections: 51.422
Number of unique NFT: 3.651.730

Most Active Sellers/Buyers*:* We identified the most active sellers and buyers in the dataset by looking at their transaction counts. This allowed us to pinpoint key participants in the market and observe trends in user behavior.

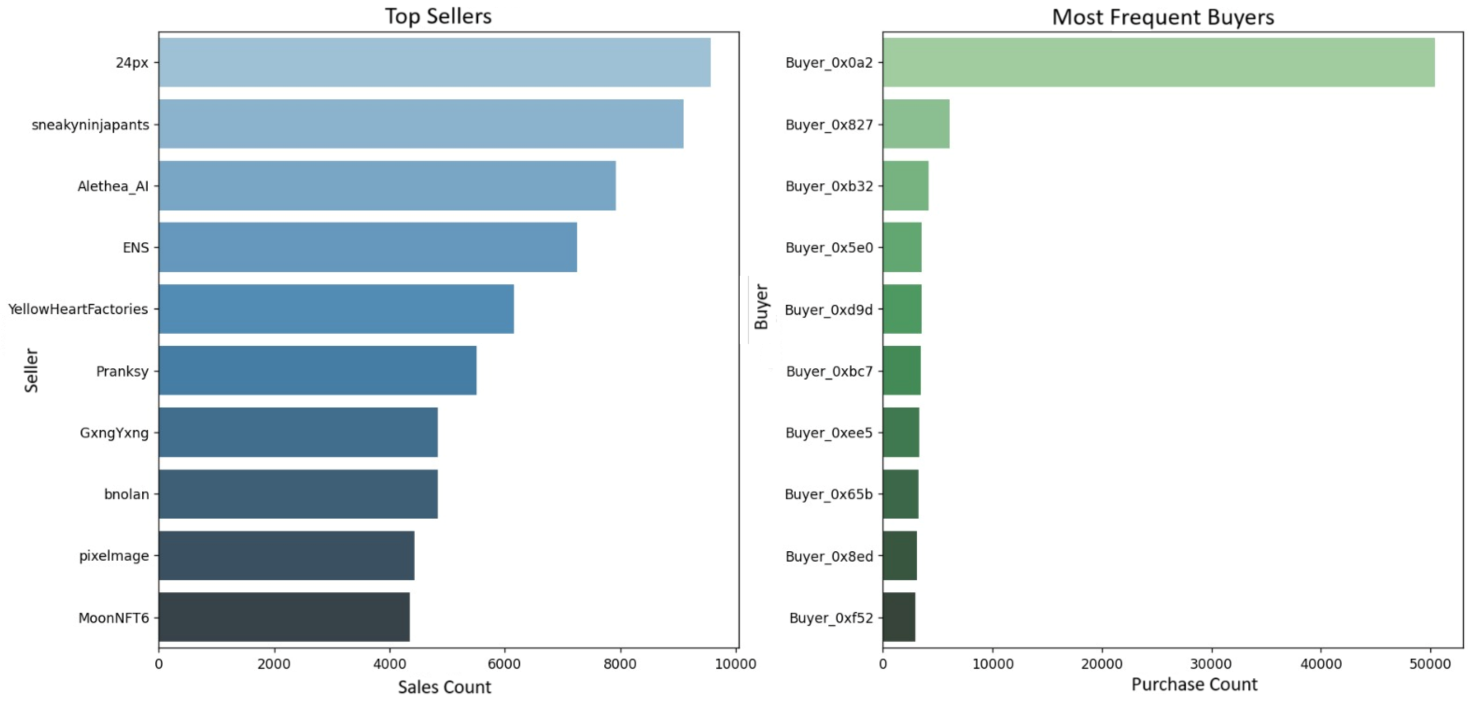


Fig. 4. Tables showing the number of sales and purchases of the most active sellers and buyers.

Sales Activity Overview: We calculated the average number of sales per user and total number of sellers to understand the typical activity level of sellers on the platform.

Average number of sales per seller: 19.27

Total number of sellers: 242.257

5. Conclusion

In this progress report, we have detailed the preliminary steps taken in our analysis of the NFT market using a comprehensive dataset containing over 5.2 million records. Our work began with an initial exploration phase, where we sought to understand the structure and characteristics of the dataset. We performed data cleaning procedures by examining missing values and assessing attribute completeness. For example, the asset.collection.short_description column was found to be missing in over 99% of the entries and was therefore removed to maintain the integrity of the analysis. Other columns with lower rates of missing data were retained and handled accordingly.

We also categorized variables into numerical and categorical types, enabling us to apply appropriate statistical and frequency-based methods. Through our basic statistical analysis of numerical attributes, we gained insight into the overall distribution, spread, and outliers within the data. For instance, the payment_token.usd_price column revealed a large standard deviation, indicating notable variability in transaction values. The frequency analysis of categorical attributes, particularly the payment_token.name column, showed a heavy concentration around a few dominant tokens such as Ether and Wrapped Ether, with Ether alone accounting for over 94% of all transactions. This suggests a strong centralization of trading activity around a limited set of cryptocurrencies, a phenomenon also observed in prior research on NFT marketplaces [11].

Moreover, we identified key participants in the market by ranking the most active buyers and sellers based on transaction counts. This step is crucial for understanding market dynamics and identifying potentially anomalous behavior, such as repeated transactions by the same entities which may suggest wash trading—a known manipulation tactic in NFT marketplaces [12]. We also analyzed the overall sales activity, calculating the average number of sales per seller, which was found to be approximately 19.27 across a pool of over 240,000 distinct sellers.

Looking forward, our analysis will shift toward advanced analytical techniques. Specifically, we aim to:

- Implement time-series and clustering algorithms to detect market trends and pricing anomalies.
- Examine feature importance to identify key factors that influence NFT prices, such as collection popularity, historical sales, or asset metadata.
- Investigate market manipulation patterns using network analysis and anomaly detection algorithms to identify suspicious trading loops or unusual transaction clusters.
- Enrich the study with visual analytics, allowing stakeholders to interact with and interpret complex patterns within the NFT ecosystem.

These steps are aligned with existing literature in blockchain data science, where a combination of statistical, machine learning, and graph-based approaches has been effective in uncovering latent structures in decentralized digital markets [13]. Our ultimate goal is to provide data-driven insights that contribute to a more transparent, trustworthy, and economically sound NFT ecosystem.

6.References

1. CryptoPotato. (2023, January 27). *Nearly 60% of NFT trading volumes in 2022 was wash trading: Report*. https://cryptopotato.com/nearly-60-of-nft-trading-volumes-in-2022-was-wash-trading-report/

2. Yang, L., Liu, Y., Xu, W., & Shi, Y. (2025). Wash trading in NFT markets: Evidence, detection, and implications. *Journal of Financial Innovation, 11*(1), Article 12. https://doi.org/10.1186/s40854-025-00766-z

3. Wong, B. (2021). *OpenSea NFT sales 2019–2021* [Data set]. Kaggle. https://www.kaggle.com/datasets/bryanw26/opensea-nft-sales-2019-2021/data

4. Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: Forecasting and control* (5th ed.). Wiley.

5. Benesty, J., Chen, J., Huang, Y., & Cohen, I. (2009). Pearson correlation coefficient. In *Noise reduction in speech processing* (pp. 1–4). Springer. https://doi.org/10.1007/978-3-642-00296-0_5

6. MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1, pp. 281–297). University of California Press.

7. Han, J., Pei, J., & Yin, Y. (2000). Mining frequent patterns without candidate generation. *ACM SIGMOD Record, 29*(2), 1–12. https://doi.org/10.1145/335191.335372

8. Chen, S., Wang, L., & Zhang, H. (2023). Detecting anomalous trading in NFT markets using network analytics. *Journal of Digital Finance, 9*(2), 113–130.

9. Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining* (pp. 413–422). IEEE. https://doi.org/10.1109/ICDM.2008.17

10. Cleveland, R. B., Cleveland, W. S., McRae, J. E., & Terpenning, I. (1990). STL: A seasonal-trend decomposition procedure based on loess. *Journal of Official Statistics, 6*(1), 3–73.

11. Dowling, M. (2022). Fertile LAND: Pricing non-fungible tokens. *Finance Research Letters, 44*, 102096. https://doi.org/10.1016/j.frl.2021.102096

12. von Wachter, V., Jensen, J. A., & Schüritz, R. (2022). NFT wash trading: Quantifying suspicious behaviour in NFT markets. *arXiv preprint* arXiv:2202.05844.

13. Regner, F., Urbach, N., & Schweizer, A. (2019). NFTs in practice–Non-fungible tokens as core component of a blockchain-based event ticketing application. In *Proceedings of the 40th International Conference on Information Systems (ICIS 2019).*