An Empirical Data Mining Study on Non-fungible Token (NFT) Markets

by

Karly (Karlygash) Kussainova

B.Sc., University of Central Asia, 2021

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Science

in the
School of Computing Science
Faculty of Applied Sciences

SIMON FRASER UNIVERSITY
Summer 2023

Declaration of Committee

Name: **Karly (Karlygash) Kussainova**

Degree: **Master of Science**

Thesis title: **An Empirical Data Mining Study on Non-fungible Token (NFT) Markets**

Committee: **Chair:** Jiangchuan Liu
Professor, Computing Science

Jian Pei
Supervisor
Professor, Computing Science

Tianzheng Wang
Committee Member
Assistant Professor, Computing Science

Jiannan Wang
Examiner
Associate Professor, Computing Science

Abstract

Non-fungible tokens (NFTs) have garnered significant attention as a unique asset class, but comprehending their pricing and making well-informed investment choices remains challenging. To tackle this challenge, we conducted an extensive data mining analysis of the NFT markets, examining transaction frequency, category preferences, price distributions, and inequality. Employing hierarchical clustering, we organized NFTs into distinct clusters based on their sales history and price dynamics. We developed predictive models for estimating NFT prices, utilizing linear regression, regularization techniques, and the Multi-Layer Perceptron (MLP) model. This research yields valuable insights into the NFT markets, empowering investors and artists to make informed decisions. Furthermore, it contributes to the broader field of digital assets, promoting market fairness and transparency. By comprehending the factors influencing NFT prices and organizing NFTs into clusters, this study enhances transparency and facilitates equitable valuation.

Keywords: NFTs, NFT market, statistical analysis, hierarchical clustering, regression, Multi-Layer Perceptron (MLP), transparency, fairness

Dedication

I would like to dedicate this thesis to the following individuals who have played instrumental roles in my academic journey:

To my esteemed supervisor, Jian Pei, I am deeply grateful for your unwavering support, guidance, and mentorship throughout this entire process. Your expertise, patience, and timely feedback have been invaluable in shaping my research and helping me overcome various challenges. Despite the time difference, you have consistently been there to provide academic and non-academic assistance, and for that, I am truly thankful.

To my loving family, your unwavering belief in me and continuous encouragement have been a constant source of inspiration. Your emotional support and understanding during the highs and lows of this journey have helped me stay motivated and focused. I am grateful for your presence in my life and for reminding me to never give up, even when the path seemed arduous.

To my partner, Soroush, your continuous support and understanding have been the pillars of strength throughout this entire program. Your love, care, and selflessness in taking on additional responsibilities when I was occupied with studies and work have been instrumental in my success. Your belief in me and your constant encouragement have kept me going even in the most challenging times.

To my dedicated labmates, your camaraderie, shared knowledge, and collective pursuit of excellence have made this journey more meaningful and enjoyable. Your support in navigating through the academic demands, course loads, and thesis writing has been invaluable. I am grateful for the collaborative environment we have fostered, as it has contributed to my growth and achievement.

This thesis is a testament to the collective support and encouragement I have received from these remarkable individuals. Their belief in me and their unwavering presence in my life have been instrumental in my accomplishments. I dedicate this work to them with sincere appreciation and heartfelt gratitude.

Acknowledgements

First and foremost, I would like to extend my heartfelt thanks to my supervisor, Jian Pei. Your guidance, expertise, and unwavering support have been instrumental in shaping my research and academic growth. Your valuable feedback and timely assistance, despite the time difference, have been invaluable throughout my entire journey.

I would also like to acknowledge the faculty members at [University/Institution Name] for their flexibility and care. Their dedication to providing a nurturing academic environment has played a vital role in my development as a researcher and scholar.

I am grateful to Dr. Angelica Lim for her invaluable support during challenging times. Her encouragement, empathy, and words of wisdom have been a constant source of inspiration and motivation.

I would like to extend my appreciation to my committee members for their valuable insights and for dedicating their time to review and provide feedback on my work. Your guidance and constructive criticism have greatly contributed to the improvement of this thesis.

Additionally, I would like to recognize the contributions of two members of the SFU team who have had a significant impact on my success. Dr. Steven Bergner from SFU Big Data Hub has been a valuable collaborator and mentor, providing me with guidance and support beyond the classroom. I am also grateful to Eva Yap from SFU Outreach, with whom I have had the opportunity to work on meaningful projects outside of academia. Your support and involvement have broadened my horizons and enriched my overall learning experience.

Finally, I want to express my deep appreciation to my family and friends for their unwavering support, understanding, and encouragement throughout this journey. Their belief in me, even during moments of self-doubt, has been a constant source of strength.

To everyone mentioned above, as well as those whose names may not be listed but have contributed to my academic and personal growth in various ways, please accept my sincerest gratitude. Your support, guidance, and belief in my abilities have been invaluable, and I am truly grateful for the impact you have made on my life.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

Non-fungible tokens (NFTs) have become increasingly popular as a new asset class with potential for investment and speculation. In recent years, the NFT market has experienced explosive growth, with millions of dollars spent on individual NFTs [37]. This has led to a growing interest in understanding the factors that influence the prices of NFTs and how they can be used as potential investment opportunities.

NFTs are unique digital assets that can represent ownership of various items, such as artwork, music, and collectibles [4]. They are created using blockchain technology, which provides a secure and transparent way to verify ownership and authenticity [38]. Unlike traditional assets, NFTs are indivisible and cannot be replicated, making them scarce and potentially valuable. The process of creating and trading NFTs involves a complex ecosystem of platforms, wallets, and marketplaces, each with its own set of rules and features.

Another definition of NFTs states that they are digital assets that feature identifying information documented in smart contracts that can be defined as digital contracts allowing terms contingent on a decentralized consensus that is tamper-proof and typically self-enforcing through automated execution [11]. NFTs are cryptographically secured digital assets, usually called digital tokens [23], on a blockchain network that provides a representation of a unique item. The first NFT prototype was Colored Coins on the Bitcoin blockchain [35]. Colored Coins allowed for further experimentation and created the prerequisites for the birth of NFTs on the Ethereum blockchain later in 2016.

The uniqueness and scarcity of NFTs have led to some high-profile sales in recent years. For example, in March 2021, a digital artwork by the artist Beeple sold for $69 million at a Christie's auction [33]. Other notable NFT trades include a tweet by Twitter CEO Jack Dorsey, which sold for $2.9 million, and a virtual house in the online game Decentraland, which sold for $2.4 million [20] [22]. These sales highlight the potential for NFTs as a new form of valuable asset.

The NFT market is influenced by various factors, such as the overall health of the cryptocurrency market, changes in regulation and legal frameworks, and the emergence of new platforms for NFT creation and trade. Some experts suggest that the success or adoption of younger NFT projects is also influenced by that of more established markets such as the cryptocurrency market [7]. These factors can all play a role in determining the value and appeal of NFTs, making it important to understand and analyze them to gain insights into the NFT market and its potential for investment and growth.

1.2 Motivation

The first motivation for studying NFT sales is to gain a deeper understanding of the factors that influence their prices. The NFT market is relatively new and lacks traditional valuation models, which makes it difficult to determine how prices are set. By analyzing the transaction data and using statistical and machine learning techniques, we hope to identify the key factors that affect NFT prices and develop accurate prediction models. This information can be valuable not only to investors but also to artists and creators who want to understand the value of their work in the NFT market. Additionally, it can provide valuable insights into the broader art market, which is shifting towards digital assets like NFTs.

The second motivation for studying NFT sales is to promote transparency and fairness in the market. Investing in NFTs can be risky due to the lack of regulation and transparency in the market. Without proper information and guidance, investors may fall prey to scams, fraud, or misinformation. By providing insights into the pricing mechanisms and market trends, we can help investors make more informed decisions and prevent arbitrage. This can lead to a more efficient market that benefits both creators and investors. Furthermore, understanding the value of NFTs and the factors that influence them can lead to fair intellectual property and prevent fraud and exploitation. This can encourage more artists and creators to participate in the NFT market, knowing that their work is protected and valued appropriately.

1.3 Research Objectives and Questions

This research aims to analyze an open-source dataset containing NFT transaction data from 2017 to 2021 [29], with the goal of understanding the factors that influence NFT prices and developing accurate prediction models. The research questions we will address include:

- How do statistical analysis techniques contribute to our understanding of the factors that influence NFT prices, and what are the key insights gained from this analysis?
- What are the clusters identified through hierarchical clustering techniques, and what can these clusters tell us about the characteristics of different NFT categories and their respective price ranges?

- Can machine learning models accurately predict NFT prices, and what are the most important features or variables in these models that contribute to accurate price prediction?

1.4 Methodology and Dataset

We use an open-source dataset containing detailed information on NFT transactions, including unique identifiers, prices, categories, and transaction dates. We will employ statistical analysis, hierarchy of clusters, and machine learning techniques, such as regression models and Multi-Layer Perceptrons (MLPs), to analyze the dataset and address the research questions.

For the initial phase of the analysis, we utilized statistical analysis techniques to investigate the dataset. Descriptive statistics were employed to examine the fundamental characteristics of the data, including counts, missing values, unique values, and data types, as well as to identify the types of transactions present in the dataset. We then proceeded to investigate the distribution of prices and transaction volume across various categories. In addition, we adopted economic methods such as the Gini Index and Lorenz Curve to analyze the price distribution. Furthermore, we performed time series analysis and correlation analysis to identify any trends and relationships within the dataset.

The second part of the research involved applying K-means clustering to identify the main clusters from the NFT dataset. To determine the optimal number of clusters, the elbow method was used. The main characteristics of each cluster were analyzed, and clusters larger than 40% were further clustered from within, allowing us to build the hierarchy of clusters. This method of analysis enabled us to gain a better understanding of the characteristics of each cluster and the overall structure of the NFT market.

In the final stage of the research, we applied linear regression and deep learning techniques to predict NFT prices based on a selected set of features. Regularization techniques, including Lasso (L1) and Ridge (L2) regression, were implemented in linear regression models to address the issue of overfitting. The performance of the models was evaluated based on R^2 score, root mean squared error (RMSE) estimate, and alpha hyperparameter for regularization. For deep learning, we used the Multilayer Perceptron (MLP) as a regression model for price prediction. We conducted experiments by varying the batch size, epochs, activation functions, and optimizer to find the optimal combination. The performance of the model was evaluated based on R^2 score and validation loss.

1.5 Significance

Overall, this study is significant in providing a deeper understanding of the NFT market and its potential as a new asset class. The results of this research can have implications not only for investors but also for artists, collectors, and other stakeholders in the NFT ecosystem.

Chapter 2

Related Work

Previous research on NFTs has focused on various aspects of this emerging market, including the factors influencing prices, the impact on the creative economy, and the technical aspects of NFT creation and trade along with the evolution of the market. In this section, we aim to analyze the current state and research related to the NFT market from different perspectives. We will start with describing technical components of NFTs and the market development and its pricing determinants. Next, we will cover the analysis of the NFT market along with different prediction models that were implemented by researchers from around the world.

In their report [36] Qin Wang et al. provide a systematic study of the current NFT ecosystems and explore various aspects, including technical components, protocols, standards, security evaluation, opportunities, and challenges. They identify the core technical components (blockchain, smart contracts, address and transaction, data encoding) used to construct NFTs and present their protocols, standards, and targeted properties. The report [36] explains two design patterns for establishing the Non-Fungible Token (NFT) paradigm: Top to Bottom and Bottom to Top that include NFT owner and NFT buyer roles. In the Top to Bottom design, the process involves digitizing the raw data, storing it in an external database or blockchain, signing a transaction, and finally minting and trading the NFT. On the other hand, in the Bottom to Top design, the founder initiates a template, and the buyer customizes the NFT product by adding additional features on top of basic lines. The process continues with minting, trading, and finally persistently storing the NFT on-chain after the consensus procedure is completed. Each NFT has a unique identifier called a "tokenId," which is a unit256 variable. The combination of the contract address and tokenId is globally unique, and tokenId can be used as an input to generate special identifications such as images in the form of zombies or cartoon characters. Moreover, the key properties of NFT schemes as decentralized applications include verifiability, transparent execution, availability, tamper-resistance, usability, atomicity, and tradability.

The technical report summarized above provides us with basic understanding of how NFTs are being traded and highlights the importance of continuing research in this area as it can facilitate the boosting in the gaming industry, flourishing of the virtual events, protection of the digital collectibles, and inspiring the Metaverse [36].

Kräussl and Tugnetti [25] describe the development of the NFT market in the four stages. The first stage began in 2012 with the Colored Coins prototype on the Bitcoin blockchain, and in 2017, the ERC-721 standard was proposed for NFTs on the Ethereum blockchain. The second stage started with the success of NFT collections like CryptoKitties and CryptoPunks in 2017. The third stage began in 2018 when some platforms decided to create their own blockchain to offer users an alternative to Ethereum. The fourth stage began in 2020 when NFTs gained popularity among the general population, attracting investors and artists to create their own series. In their review [25], five major categories of the NFTs were identified as follows: gaming, collectibles, utilities, art, and metaverse. Moreover, the authors were also able to identify the five major approaches applied to the determination of pricing NFTs: hedonic regression models (estimates the relationship between the price of a good and its intrinsic characteristics), repeat sales regressions (examines the change in the value of an asset over time), vector autoregressive models (analyzes the relationship between multiple time series), machine learning, and wavelet models (analyzes signals and time series data at multiple scales) [25].

The review conducted above has highlighted the significant events and milestones that have shaped the evolution of the NFT market, providing valuable insights into the factors that could affect pricing strategies [25]. In addition to the NFT categories mentioned in previous studies, we have introduced an additional category, 'other', to capture a wider range of NFT types. Building on the pricing determination techniques discussed in the literature, our approach to NFT price prediction combines repeat sales regressions and machine learning methods to enable accurate and reliable price forecasting based on time-series data and sales price information.

Predicting the sale price was also attempted by another group of researchers, where they analyzed the NFT market using data from the OpenSea platform, exploring transaction volumes, prices, trade networks, and visual features that contribute to the value of NFTs [28]. The authors identify several factors that influence NFT prices, including the rarity of the artwork, the reputation of the artist, and the aesthetic value of the NFT. They also use computer vision and dimensionality reduction techniques to analyze the visual features of NFTs, such as color, shape, and composition, finding that certain features, like bright colors and symmetry, are associated with higher NFT prices.

Despite the fact that both of our research use the same dataset, there are some similarities and differences. Like the authors in [28], we also analyzed transaction volumes and prices of NFTs to understand the factors that influence their prices. However, our research goes beyond by using additional statistical methods like the Gini index and Lorenz curve analysis to investigate the distribution of wealth in the NFT market. Furthermore, we employed machine learning techniques, including linear regression and multilayer perceptron models, to predict NFT prices with a higher level of accuracy. Moreover, while the authors in [28] focused on the visual features of NFTs, our research investigated the impact of NFT categories and explored patterns in the data using hierarchical clustering techniques. We also analyzed the transaction history of NFTs, including the number of times they were sold and the length of time between sales, to gain a better understanding of how NFT prices evolve over time.

In another research, Costa, La Cava, and Tagarelli address the problem of predicting the financial performance of Non-Fungible Tokens (NFTs) using their associated images and textual descriptions [13]. They propose a novel multimodal deep learning framework called MERLIN, which uses Transformer-based language and visual models and graph neural network models to learn dense representations of NFT data. The objective of the framework is to perform a price-category classification into low, medium, high, and predict the price category of a previously unseen NFT.

Similar to the current research, the authors used a publicly available dataset on NFT purchase transactions, which includes sales from various NFT markets [29]. They selected 202,257 NFTs with images and descriptions that had at least one secondary sale and defined three price categories based on the distribution of average prices. While both [13] and our research aim to predict the prices of NFTs based on the same dataset, our methodology differs in terms of the features that are being selected for training purposes. While the features selected for [13] include text and image data, the features selected for this research include prices of the NFTs in USD and corresponding dates of the sales.

Another study conducted by Gumelar et al. [18] aims to predict the price of NFT coins, which consist of various types, such as mana coins, sand, axes, and other NFT coins. These coins are required to purchase NFT assets, and as NFTs are being increasingly used for trading in the blockchain, the movement of NFT coins is erratic and uncertain, making it difficult for investors to make profitable decisions. Through the literature study stage, interviews, and viewing daily NFT coin price data, where the attributes used are date, open, high, low, close, and volume, the authors [18] were able to apply k-Nearest Neighbors (KNN) algorithm for price prediction. The study collects data from coinmarketcap.com for the period January 1, 2019, to December 31, 2021, and processes it to predict the close value of the NFT coin price for a period of one day.

Although both the current study and the research summarized above [18] share the one of the same objectives of helping investors make profitable decisions using machine learning techniques, we focus on the price prediction of the NFT asset, while Gumelar et al. focus on the trade sales of the NFT coins.

Chapter 3

Preliminaries: Key Concepts and Statistical Methods

3.1 Gini Index and Lorenz Curve

The Gini Index and Lorenz Curve are important statistical tools for analyzing the distribution of NFT values. The Lorenz Curve [27] is a graphical representation of the distribution of a particular variable, with the cumulative percentage of the population on the x-axis and the cumulative percentage of the variable on the y-axis. It is named after the American economist Max O. Lorenz, who first used it to study income inequality in 1905 [9]. Suppose that some quantity Q, which could stand for wealth, income, family income, land, food, and so on, is distributed in a population. If we imagine the population to be lined up by increasing the order of their shares of Q, then for any p between 0 and 1 the people in the first fraction p of the line represent the Q-poorest of the population. We then call L(p) the fraction of the totality of Q owned by that fraction of the population [15]. In the case when everyone has the exact same amount of Q, the order of our imaginary line-up would give us a perfect equitability (Fig 3.1).

The Gini Index [17] is a measure of inequality in the distribution of a particular variable, such as wealth or income. It is calculated by taking the ratio of the area between the Lorenz Curve and the line of perfect equitability (blue area -B from Fig.1.) to the total area under the Lorenz Curve [9]. A Gini Index of 0 indicates perfect equality (every individual has the same value) and a Gini Index of 1 indicates perfect inequality (one individual has all the value). The Gini index can be calculated using:

$$G = \frac{A}{A+B} \tag{3.1}$$

$$G = 1 - \frac{1}{2}\int L(p)dp \tag{3.2}$$

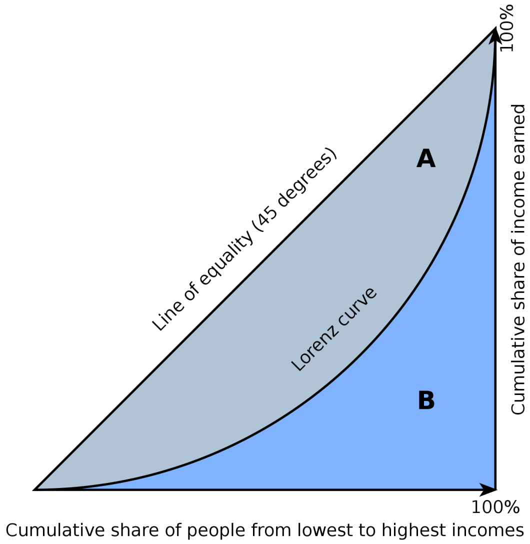

100%
A
Line of equality (45 degrees)
Lorenz curve
B
Cumulative share of income earned
100%
Cumulative share of people from lowest to highest incomes

Figure 3.1: Gini Index Visual Representation[12]

In the context of NFTs, the Gini Index and Lorenz Curve can be used to analyze the distribution of value among different types of NFTs, as well as the distribution of value among NFT holders. This is important for understanding the overall level of inequality within the NFT market and can inform policy decisions related to NFT regulation and taxation. For example, if the Gini Index is high and the Lorenz Curve is highly skewed, this could indicate that a small number of NFTs or NFT holders are capturing a disproportionate share of the value. Conversely, if the Gini Index is low and the Lorenz Curve is relatively flat, this could indicate a more equal distribution of value. Overall, the Gini Index and Lorenz Curve provide a useful framework for analyzing the distribution of NFT value and can inform decision-making within the NFT market.

3.2 Moving Averages

The moving average method [8] is a popular time series analysis method used to understand the characteristics of a time series and to find a pattern or formula to predict future events. Simple Moving Average (SMA) is the easiest type of moving average to calculate, where each point in the time series data is weighted the same regardless of its position in the sequence. Although many other soft computing methods have been developed, the moving average method is still considered as one of the best methods due to its ease, objectiveness, reliability, and usefulness [19]. For every i-th data point, at a given time t, the Simple Moving Average (SMA(t)) with n data points can be calculated as follows:

$$SMA(t) = \frac{1}{n} \sum A(i) \tag{3.3}$$

Moving average methods come in various forms, but their underlying purpose remains the same: to track the trend determination of the given time series data. Short-term fluctuations are typically 'smoothed' using a rolling window of a short time period (e.g. 3 days, 5 days, 7 days). By smoothing out short-term fluctuations, moving averages can help identify underlying trends, which can be useful for predicting future values of NFTs or making informed decisions based on the trend direction of the NFTs.

3.3 Correlation Analysis

The use of correlation analysis is an important statistical tool for identifying relationships between different variables in a dataset. In finance, correlation analysis is widely used to identify the relationship between different assets, such as stocks, bonds, and commodities. Additionally, correlation analysis can be used to identify trends in market behavior and to inform investment decisions.

One commonly used method for calculating the correlation coefficient is the Pearson correlation coefficient [32], which measures the strength and direction of the linear relationship between two variables. This coefficient ranges from -1 to 1, where a value of -1 indicates a perfectly negative correlation, a value of 1 indicates a perfectly positive correlation, and a value of 0 indicates no correlation. For given pair of variables or vectors X and Y, the formula to calculate the Pearson Correlation coefficient is as follows :

$$\rho_{XY} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y} \tag{3.4}$$

where E is an expected value, σ_X and σ_Y are the standard deviations of X and Y, and μ_X and μ_Y are the mean values of X and Y.

Pearson correlation is widely used in financial analysis and is particularly useful when analyzing time-series data, such as stock prices or market indices. According to Ezekiel [14], correlation analysis is a powerful tool that can be used to identify relationships between different variables in a dataset. The method can be used to identify patterns and trends in data, and can provide important insights into the behavior of complex systems. Additionally, correlation analysis is widely used in finance to inform investment decisions and to identify opportunities for diversification. The Pearson correlation coefficient is a widely used method for calculating correlations and is particularly useful for analyzing time-series data of the NFTs transactions.

3.4 Clustering Algorithms and the Hierarchy of Clusters

The K-means clustering algorithm [21], introduced by Hartigan and Wong in 1979, is a widely used unsupervised learning method that partitions data into K clusters based on their similarity [21]. The algorithm works by randomly selecting K centroids, which represent the center of each cluster, and then assigning each data point to the nearest centroid based on its distance. In R^D space, for data points x and x', the natural choice of distance computed is Euclidean [1]:

$$||x - x^{'}|| = \sqrt{\sum_{d=1}^{D} = (x_d - x^{'}_d)^2} \tag{3.5}$$

After the random assignment of the centroids, they are then recalculated as the mean of the data points assigned to them, and the process is repeated until convergence, where the centroids no longer change significantly. The algorithm for K-means clustering is as follows [1]:

Algorithm 1: K-means Algorithm

```
Input  : Data vectors $\{x_n\}_{n=1}^{N}$, number of clusters $K$
Output: Assignments $\{r_n\}_{n=1}^{N}$ for each data point, cluster means $\{\mu_k\}_{k=1}^{K}$
for $n \leftarrow 1$ to $N$ do
    $r_n \leftarrow [0, 0, \ldots, 0]$;
    $k' \leftarrow \text{RandomInteger}(1, K)$;
    $r_{n,k'} \leftarrow 1$;
end
repeat
    for $k \leftarrow 1$ to $K$ do
        $N_k \leftarrow \sum_{n=1}^{N} r_{n,k}$;
        $\mu_k \leftarrow \frac{1}{N_k} \sum_{n=1}^{N} r_{n,k} x_n$;
    end
    for $n \leftarrow 1$ to $N$ do
        $r_n \leftarrow [0, 0, \ldots, 0]$;
        $k' \leftarrow \arg\min_k ||x_n - \mu_k||_2$;
        $r_{n,k'} \leftarrow 1$;
    end
until none of the $r_n$ change;
```

To determine the optimal number of clusters for K-means, the "elbow method" [6] is commonly used. The idea is to determine the number of clusters then add clusters, calculate the sum squared error (SSE) or Inertia per cluster until the maximum number of clusters that have been determined, then by comparing the difference SSE of each cluster, the most extreme difference forming the angle of the elbow shows the best cluster number [7]. The formula to calculate SSE is as follows:

$$SSE = \sum_{i=1}^{k} \sum_{x_j \subset C_i} ||x_j - \mu_i|| \tag{3.6}$$

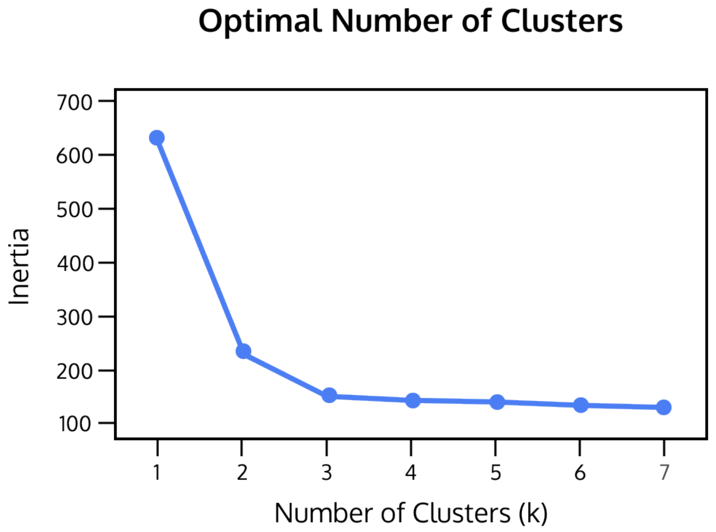


Figure 3.2: SSE (Inertia) VS number of clusters (k)[10]

where k is the number of clusters, C_i is the i-th cluster, x_j is j-th data point, and i is i-th centroid. If we visualize the results of the SSE(Inertia) calculation against the number of K (Fig 3.2):

K-means has been used in various applications, such as image segmentation, market segmentation, and anomaly detection. In finance, K-means has been used to identify groups of stocks with similar return patterns and to cluster customers based on their behavior.

Hierarchical clustering [30] is a data clustering algorithm that creates a hierarchical representation of data by recursively partitioning the dataset into smaller and smaller clusters or groups. It constructs a tree over the data: the leaves are individual data items, while the root is a single cluster that contains all of the data. Between the root and the leaves are intermediate clusters that contain subsets of the data. The main idea of hierarchical clustering is to make "clusters of clusters" going upwards to construct a tree. There are two main conceptual approaches to forming such a tree. Hierarchical agglomerative clustering (HAC) starts at the bottom, with every datum in its own singleton cluster, and merges groups together. Divisive clustering starts with all of the data in one big group and then chops it up until every datum is in its own singleton group [2]. In this research, we will use the second approach of divisive clustering to build a hierarchy of clusters using the K-means algorithm. We assume that the NFTs might be grouped into different clusters, and the groups that are extremely large might have subclusters that could be calculated using the same approach.

3.5 Regression Models and Regularization

The method of least squares [26] [16] is used as a standard approach in regression analysis, where regression models are known as a type of statistical model used to explore the relationship between an outcome variable and one or more predictor variables [24]. These models are widely used in finance, economics, and other fields to make predictions about future events or to better understand the relationship between different variables. The most common type of regression model is linear regression, which involves fitting a straight line to a set of data points in order to predict the value of the outcome variable based on the value of the predictor variable(s) [24]. Given a dataset $\{y_i, x_{i1}, ..., x_{ip}\}_{i=1}^{n}$ of n statistical units, a linear regression model assumes that there's a linear relationship between dependent variable y_i and a vector of independent variables $\{x_{i1}, ..., x_{ip}\}_{i=1}^{n}$ with an addition of noise (error value) :

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} + \epsilon = x^T \beta + \epsilon_i, \text{ for } i = 1, 2, ..., n \tag{3.7}$$

or alternatively in matrix notation,

$$y = X\beta + \epsilon \tag{3.8}$$

where x^T stands for the transpose value of x, so that x_i^T produces the inner product between vectors x_i and β; $\{0, 1, ..., p\}$ is a (p+1)-dimensional parameter vector, where β_0 is the intercept term.

When evaluating the performance of a regression model, two important metrics are often used: $R2$ score and Root Mean Squared Error (RMSE).

The R^2 score, also known as the coefficient of determination, measures how well the model fits the data by calculating the proportion of variance in the outcome variable that is explained by the predictor variable(s). The R^2 score ranges from 0 to 1, with higher values indicating a better fit. It can be calculated using:

$$R^2 = 1 - \frac{\sum(y_i - \mu_y)^2}{\sum(y_i - y')^2} \tag{3.9}$$

where y_i is the actual value of the dependent variable, μ_y is is the mean of the dependent variable, and ý is a predicted value of an independent variable.

RMSE, on the other hand, measures the error between the predicted and actual values of the outcome variable. If MSE (mean root squares) represents the average of the squared difference between the original and predicted values in the data set indicating the variance of the residuals, the RMSE measures the variance of the residuals:

$$MSE = \frac{1}{N} \sum_{i=1}^{N} (y - y^{'}) \tag{3.10}$$

and

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y - y')} \tag{3.11}$$

The goal of the regression model is to minimize RMSE and maximize the R^2 score [24].

Regularization is a technique used to prevent overfitting in regression models. Overfitting occurs when a model is too complex and begins to fit the noise in the data rather than the underlying patterns. This can lead to poor performance on new data that the model has not seen before. L1 (Lasso Regression) and L2 (Ridge Regression) regularization are two common forms of regularization used in linear regression models.

L1 regularization adds a penalty term to the cost function that is proportional to the absolute value of the model coefficients. This has the effect of shrinking some of the coefficients to zero, effectively removing them from the model. The cost function that is used to measure the error between predicted and actual values for Lasso is:

$$\sum_{i=1}^{n}(Y_i - \sum_{j=1}^{p} X_{ij})^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{3.12}$$

In this case, if λ is zero then we will get back the simple linear regression estimates whereas a very large value will make coefficients zero hence it will under-fit.

L2 regularization adds a penalty term that is proportional to the square of the model coefficients, which has the effect of shrinking all the coefficients toward zero [31]. The cost function of Ridge Regression can be computed using:

$$\sum_{i=1}^{n}(Y_i - \sum_{j=1}^{p} X_{ij})^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{3.13}$$

In this case, if λ is zero then we will get back the simple linear regression estimates whereas a very large value will add too much weight and it will lead to under-fitting.

In this research, regression models are being used to predict the prices of NFTs based on a set of historical data variables. R^2 score and RMSE will be used to evaluate the performance of the model and select the best set of input variables. In addition, we will experiment with L1 and L2 regularization to improve the performance of the model and prevent overfitting. By incorporating these techniques into the analysis, we hope to create a more accurate and reliable model for predicting NFT price values.

3.6 Multi-Layer Perceptron

Neural networks are a type of machine learning algorithm modeled on the structure of the human brain [31]. At its core, a neural network is made up of interconnected nodes, or neurons, that process and transmit information. These neurons are organized into layers,

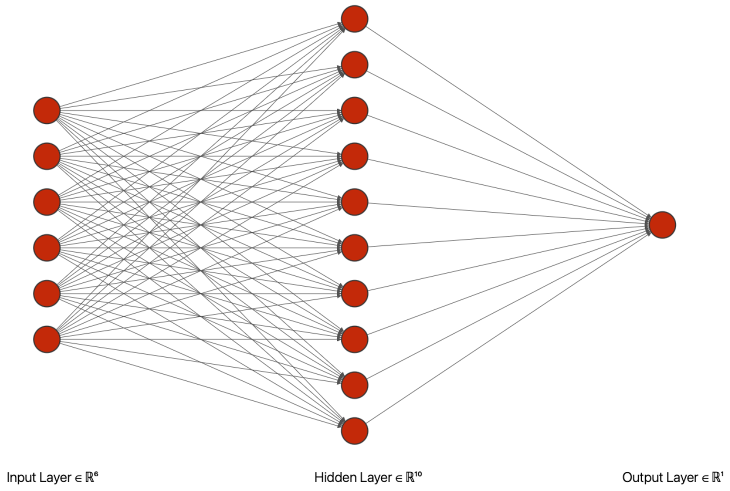

Input Layer $\in \mathbb{R}^6$
Hidden Layer $\in \mathbb{R}^{10}$
Output Layer $\in \mathbb{R}^1$

Figure 3.3: MLP structure

with each layer performing a specific function in the network. The input layer receives the raw data, while the output layer produces the final output, which could be a prediction or classification [3]. The layers in between are known as hidden layers and are responsible for processing the input data through a series of non-linear transformations to extract relevant features. Neural networks use a process called backpropagation to adjust the weights and biases of each neuron in the network during training, in order to minimize the difference between the predicted output and the actual output [31].

The multilayer perceptron (MLP) [34] is a type of neural network architecture that has been widely used for predicting outcomes based on input variables. It is a feedforward network, meaning the information flows in one direction from the input layer to the output layer, and it has one or more hidden layers in between. Each neuron in the network is connected to every neuron in the adjacent layers, and each connection has a weight associated with it (see Fig 3.3).

The MLP uses activation functions to introduce non-linearities into the network, allowing it to model complex relationships between the input and output variables. The primary role of the Activation Function is to transform the summed weighted input from the node into an output value to be fed to the next hidden layer or as output. During training, the weights and biases are adjusted using backpropagation and optimizer algorithms to minimize the difference between the predicted and actual values [3].

As weight, input, and bias enter the node, they go through a transformation of an activation functions (could be ReLu, Sigmoid, or Tanh):

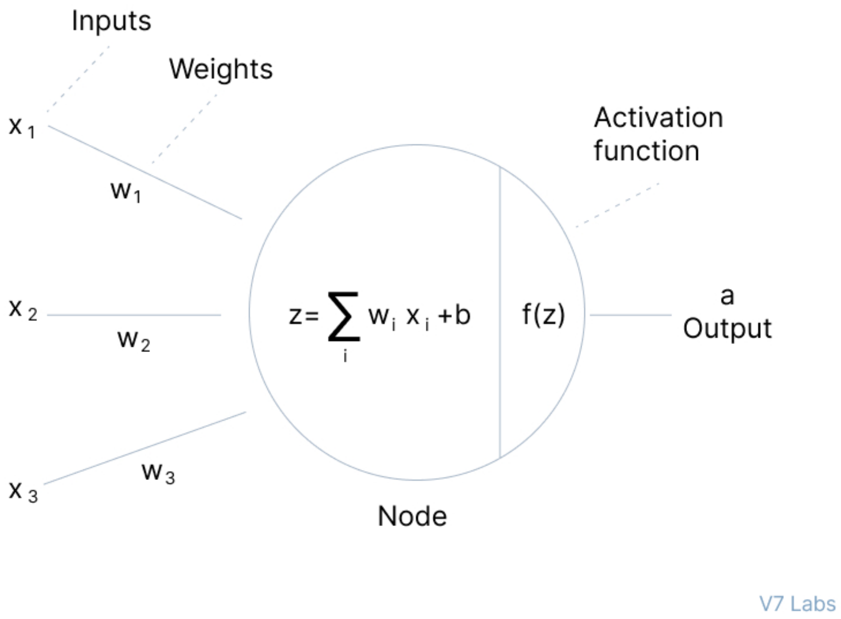


Figure 3.4: Neuron composition [5]

$$a = \phi(\sum_j w_j x_j + b) \tag{3.14}$$

where, x_j are the inputs, w_j are the weights, b is bias, and ϕ is an activation function. At every hidden layer the network computation go through transformations in the form of:

$$h_i^l = \phi^l(\sum_j w_{ij}^l x_j + b_i) \tag{3.15}$$

$$y_{last} = \phi_y(\sum_j w_i^l x_j + b_{last}) \tag{3.16}$$

where l represents an l-th hidden layer, y_{last} indicates the output value with w_j weights and blast biases that go into the output layer. When training an MLP model, it's important to consider the most optimal optimizer algorithm, the number of hidden layers (for regression 1 is typically enough), the epoch number, and the batch size. These hyperparameters need to be tuned to achieve optimal performance of the model. MLP can be used for predicting NFT prices by taking various features of NFTs as inputs and the price as the output. The MLP model will be trained on historical data, and once trained, it might be used to predict the price of new NFTs based on their features.

Similar to linear regression, when evaluating the performance of the MLP model, two common metrics used are R^2 and validation loss. The R^2 score measures the proportion of variance in the outcome variable (i.e., price) that is explained by the predictor variables (i.e., NFT features). A higher R^2 score indicates a better fit of the model to the data. Validation loss, on the other hand, measures the error between the predicted and actual values of the outcome variable during validation. The goal of the MLP model is to minimize the validation loss and maximize the R^2 score, indicating a good fit of the model to the data.

Chapter 4

Results: Statistical Analysis

4.1 Dataset

This dataset was collected by a group of scientists [28] from several NFT marketplaces such as OpenSea, Atomic, Cryptokitties, Godsunchained, and Decentraland Market APIs. The data encompasses a period of almost four years, from June 23, 2017, to April 27, 2021, and includes information on 4,040,908 unique NFTs that were traded during this time period. The dataset contains 6,071,027 trades and 24 columns of information, with key attributes of interest being ID_token (unique ID identifier for each NFT), Price_USD (the price during the transaction in USD), Category (one of the categories: Art, Games, Utility, Metaverse, Collectible, and Other), and Datetime_updated (date of transaction updated to seconds).

4.2 Descriptive Analysis

This section aims to provide an overview of the key characteristics of the NFT dataset. In this section, we explore the distribution of NFT prices and volume, the frequency and duration of trades, and the most popular categories of NFTs.

As mentioned earlier, the dataset includes 6,071,027 rows (transactions) with 4,040,908 unique ID_tokens (unique NFTs). The columns chosen for this study along with their data types can be seen on Table 4.1

Table 4.1: Description of Important Columns

Column Name	**Data Type**	**Description**
ID_token	string	Unique identifier of each NFT
Price_USD	float	Price of the sale in USD
Category	string	Category of an NFT (Art, Games, Utility, Metaverse, Collectible, Other)
Datetime_updated	datetime	Updated transaction date

Even though there are 6,071,027 transactions, most of them (2,874,819) were one-time purchases (sold_once), while the rest of the transactions were sold more than once. At the

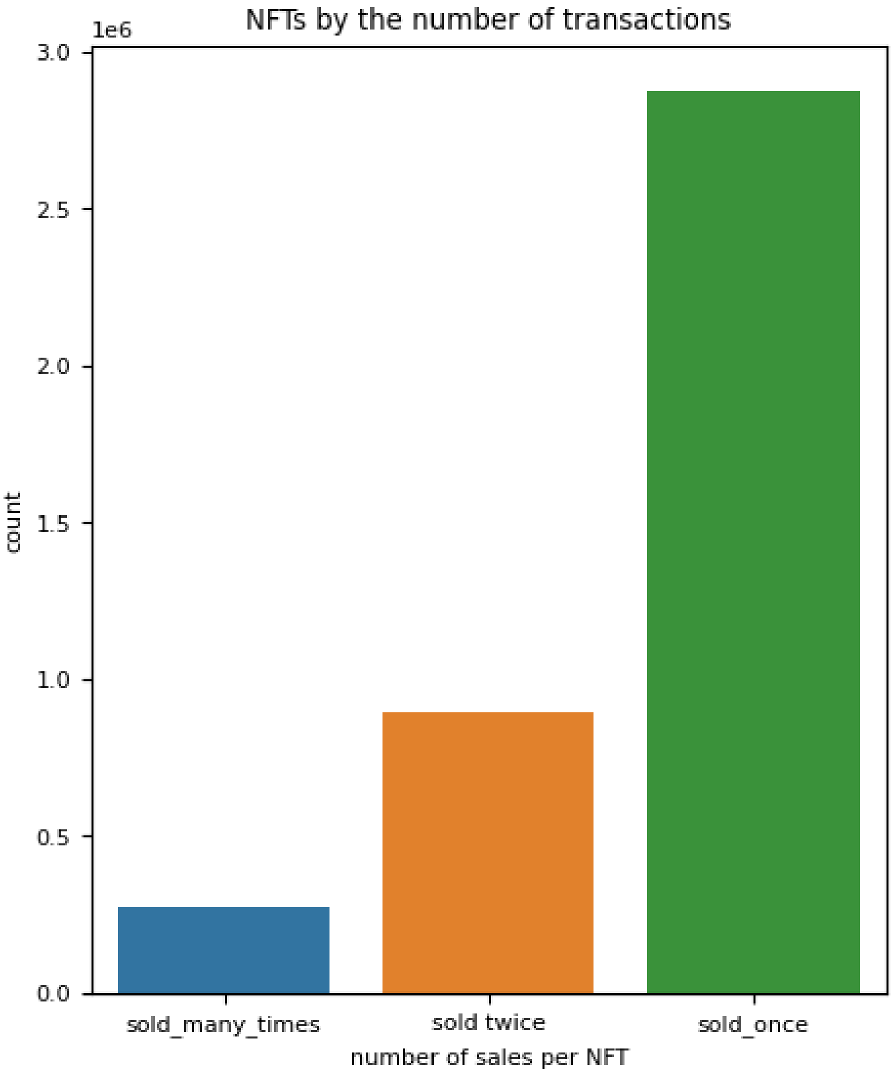


Figure 4.1: NFTs by the number of transactions

same time, the number of NFTs that were traded twice is 891,844 (sold_twice), and the remaining 274,245 NFTs were traded more than two times (sold_many_times):

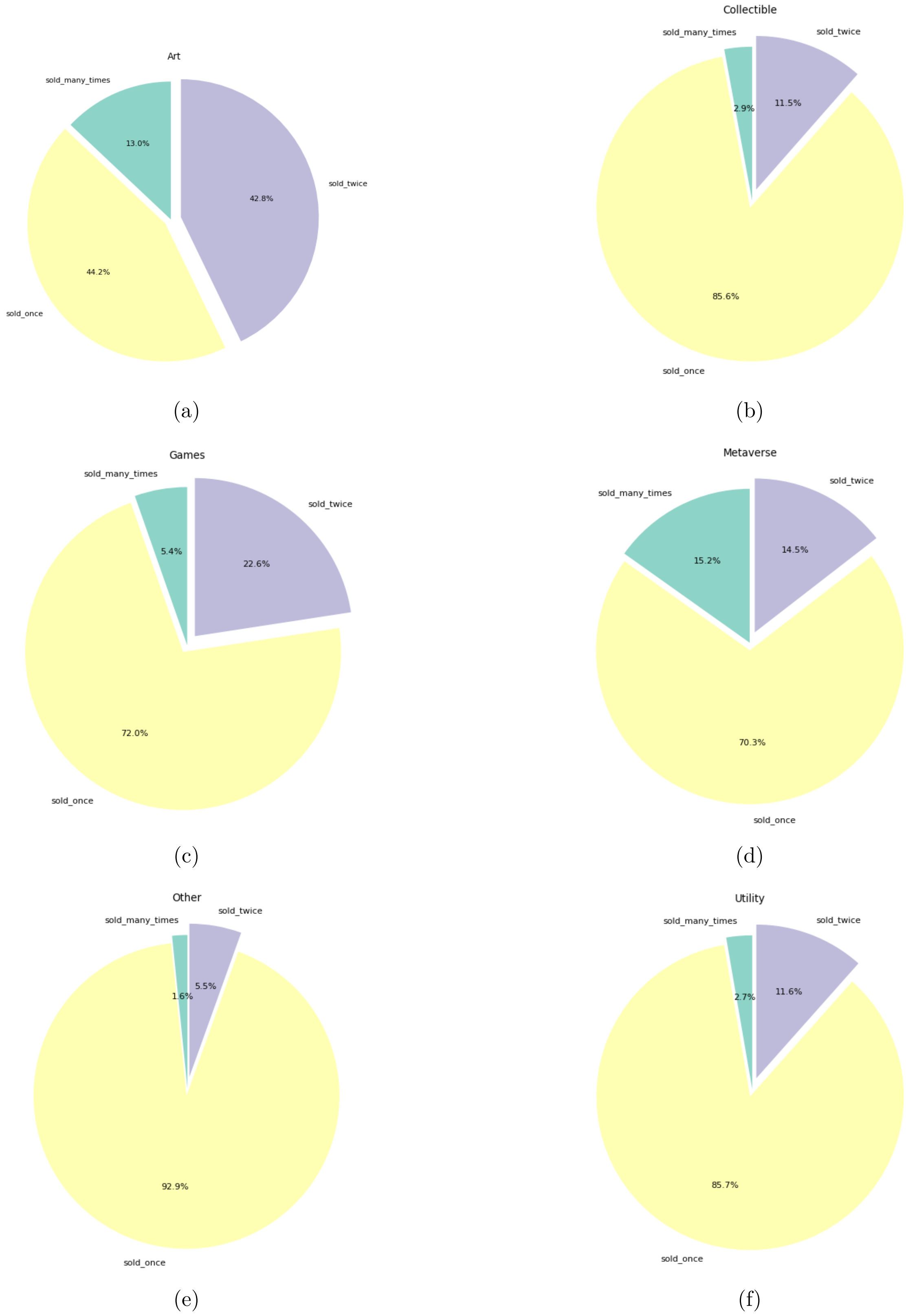
Further, applying a similar analysis to each category, we can observe that for the category of Art, there's a distinct amount of NFTs that were traded more than once (55.79%). Next, we can also see that for the categories of Metaverse and Games, there are 27.95% to 29.7% of NFTs being traded more than once, while for the rest of the categories, NFTs are mostly being traded once (over 85% of the transactions are being sold once). Even though the numbers vary from category to category, the main ratio compared to Table 4.2 is similar: the largest group is sold once, the medium group is sold twice, and the smallest group of NFTs is sold more than twice (see Table 4.3 and Fig 4.2).

Table 4.2: Sales by Type

Sale Type	**Count**	**Percent**
Sold many times	274,245	6.79%
Sold twice	891,844	22.07%
Sold once	2,874,819	71.14%

Table 4.3: Sales by Category and Type

Category	**Sale Type**	**Within Category %**
Art	sold_many_times	12.96%
	sold_once	44.21%
	sold_twice	42.83%
Collectible	sold_many_times	2.90%
	sold_once	85.63%
	sold_twice	11.47%
Games	sold_many_times	5.37%
	sold_once	72.05%
	sold_twice	22.58%
Metaverse	sold_many_times	15.17%
	sold_once	70.32%
	sold_twice	14.51%
Other	sold_many_times	1.60%
	sold_once	92.91%
	sold_twice	5.49%
Utility	sold_many_times	2.71%
	sold_once	85.73%
	sold_twice	11.56%



(a)

(b)

(c)

(d)

(e)

(f)

Figure 4.2: NFTs by Sale Type per Category

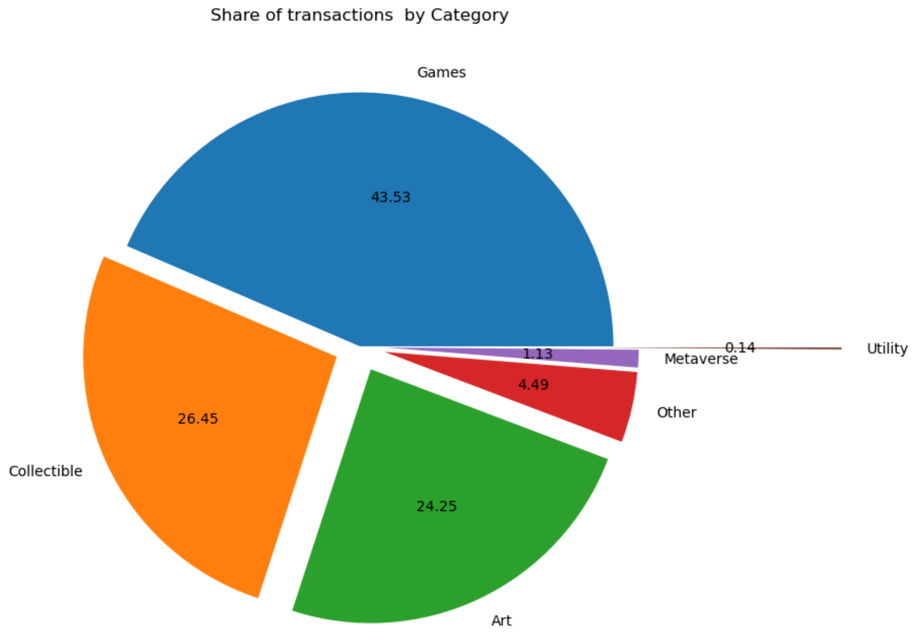


Figure 4.3: Share of Transactions by Category

There are two major ways to describe the NFTs distribution: by the number of transactions per category and by the number of NFTs per category. In the first case, we count number of transactions per each category. As can be seen from Fig 4.3, the most popular categories in terms of the number of transactions are Games (43.53%), Collectible (26.45%), and Art (24.25%) adding up to a total of 94.23%. At the same time, the remaining 5.77% is spread around Other (4.49%), Metaverse (1.13%), and Utility (0.14%) categories.

Table 4.4: Number of Transactions by Category

Category	**no_transactions**
Games	2,643,014
Collectible	1,605,657
Art	1,472,515
Other	272,772
Metaverse	68,372
Utility	8,697

In terms of the distribution of the NFTs by belonging to a certain category, the proportions are similar to the share of transactions (see Fig 4.4), where Games (45.50%), Collectible (31.45%), and Art (17.57%) make up 94.52%, while Other (4.58%), Metaverse (0.73%), and Utility (0.17%) make up 5.48%.

The price distribution of the NFT sales in this dataset ranges between $1.51*10^{16}$(minimum price) and $7.50 * 10^6$ USD (maximum price), where the cheapest NFT was sold in the cate-

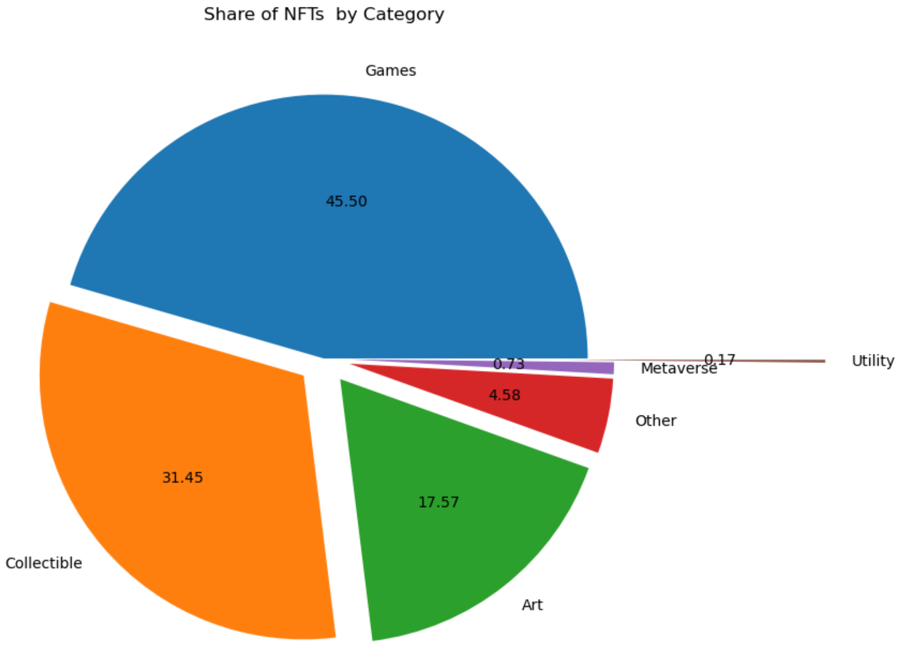


Figure 4.4: Share of NFTs by Category

Table 4.5: Number of NFTs by Category

Category	**Number of NFTs**
Games	1,919,149
Collectible	1,326,549
Art	740,892
Other	193,077
Metaverse	30,700
Utility	7,369

gory of Games of Godsunchained collection on November 22nd, 2019 and the most expensive NFT was sold in the Art category from the Cryptopunks collection on March 11th, 2021. The average price of the NFTs across all categories is 146 USD, which is not the best representation of the data due to outliers. On the other hand, the median value (50th percentile) of the prices in USD is 1.46. Such a significant difference between the mean and median can be explained by a relatively large standard deviation of 5490 USD.

Table 4.6: Statistical Measures of Price_USD

Statistical Measure	**Price_USD**
count	6.062744e+06
mean	1.463413e+02
std	5.488389e+03
min	1.514250e-16
25%	2.277600e-01
50%	1.426470e+00
75%	1.385100e+01
max	7.501893e+06

Applying the same principle to different categories of NFTs, we can see how different statistical measures appear for each category. Similar to Table 4.6 the mean value of each category is significantly higher than the median value and 75th percentile, while the standard deviation of each category is larger than the mean value, indicating the skewed distribution of the prices across the dataset.

Table 4.7: Summary Statistics by Category

Category	**Mean**	**Std**	**Min**	**25%**	**50%**	**75%**	**Max**
Art	413.237	10582.696	2.265e-16	1.090	5.662	34.415	7.501893e+06
Collectible	68.439	706.914	9.558e-16	0.546	2.091	22.050	4.291656e+05
Games	26.797	842.789	1.514e-16	0.088	0.350	3.702	5.149742e+05
Metaverse	1000.340	14740.809	1.619e-04	1.628	54.296	517.290	2.684347e+06
Other	81.553	571.256	3.831e-10	0.189	1.914	19.209	9.984975e+04
Utility	1005.444	3919.342	1.301e-10	6.387	104.927	666.222	1.303435e+05

Investigating the price distribution further, we realized that the majority of NFTs have smaller intervals, and visualizing a boxplot to see the overall distribution is not reliable. As one can see from Fig 4.5 - 4.6, most of the NFTs are populated closer to the left side of the boxplots, where price distribution is measured in USD.

To get a more clear picture, we used the 75th percentile as a threshold to see how prices are distributed across different categories. Looking at Fig 4.7, we can see that Art and Utility have the largest range of prices, Other, Collectible, and Metaverse have a similar price distribution, and the category of Games has the smallest range of prices.

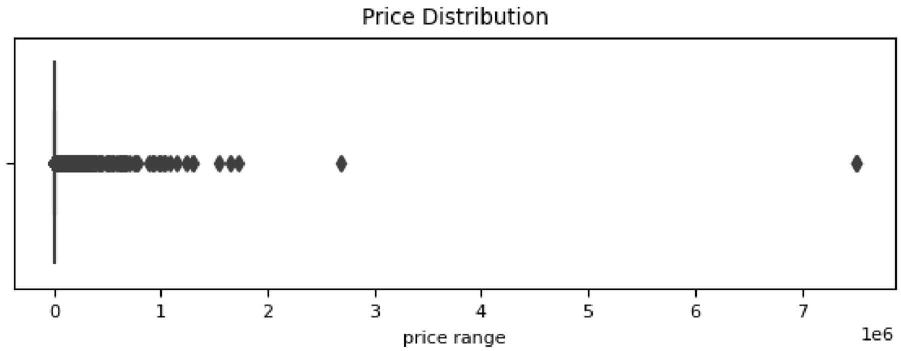


Figure 4.5: Price Distribution

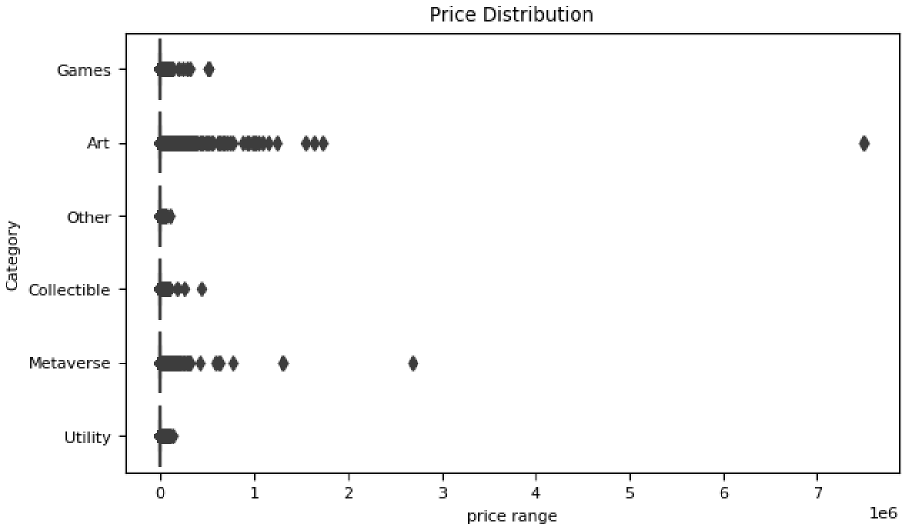


Figure 4.6: Price Distribution by Category (Boxplot)

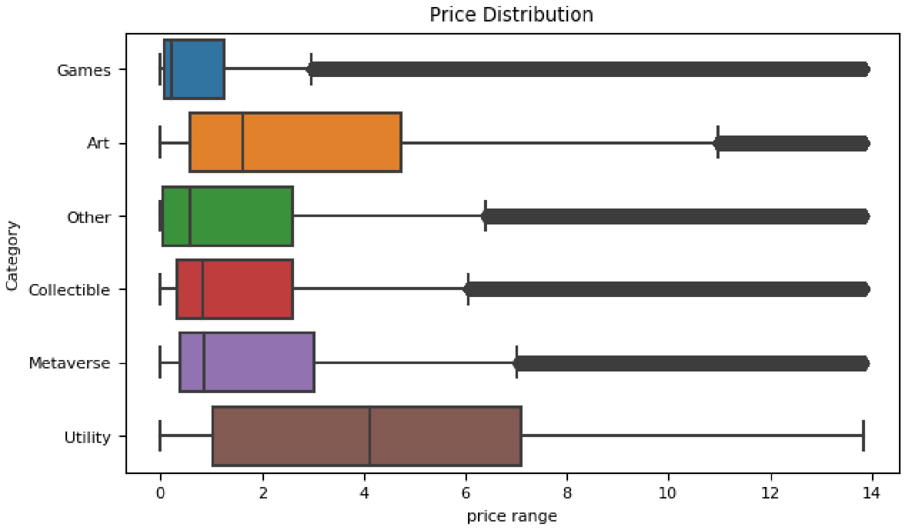


Figure 4.7: Price Distribution by Category (Limited)

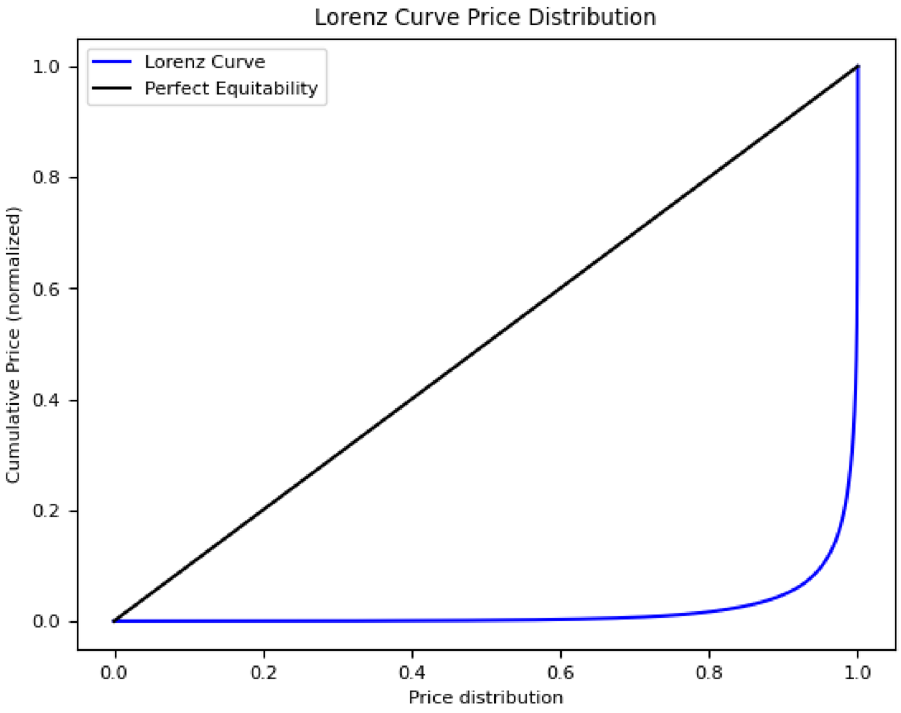


Figure 4.8: Lorenz Curve (All categories)

4.3 Gini Index and Lorenz Curve

As per the discussion in Section 3.1 the Gini index and Lorenz curve are two widely used measures to assess the inequality of a distribution. In the context of NFT prices, these measures can help us understand how the prices are distributed among different categories, and whether some categories have a higher concentration of high-priced NFTs than others. The motivation behind using the Gini index and Lorenz curve in our analysis is to provide a comprehensive picture of the distribution of NFT prices and the inequality among different categories. Even though Gini Index is known to be used for assessing the distribution of wealth in economics or other markets like real estate, here we are using it to assess the price distribution of NFTs.

As can be observed from Fig 4.8, the Lorenz Curve distribution is significantly below the perfect equitability line, which suggests that the NFT price distribution is highly unequal. Since the Lorenz Curve remains constant all the way to the end of the x-axis (0.8), a large percentage of the NFTs have a very low share of the total wealth. The curve also suddenly grows towards the end of the x-axis, which suggests that a small percentage of the population (NFT traders) holds a disproportionately large share of the total wealth.

A similar trend can be observed for all the categories with slight differences across them (see Fig 4.9)

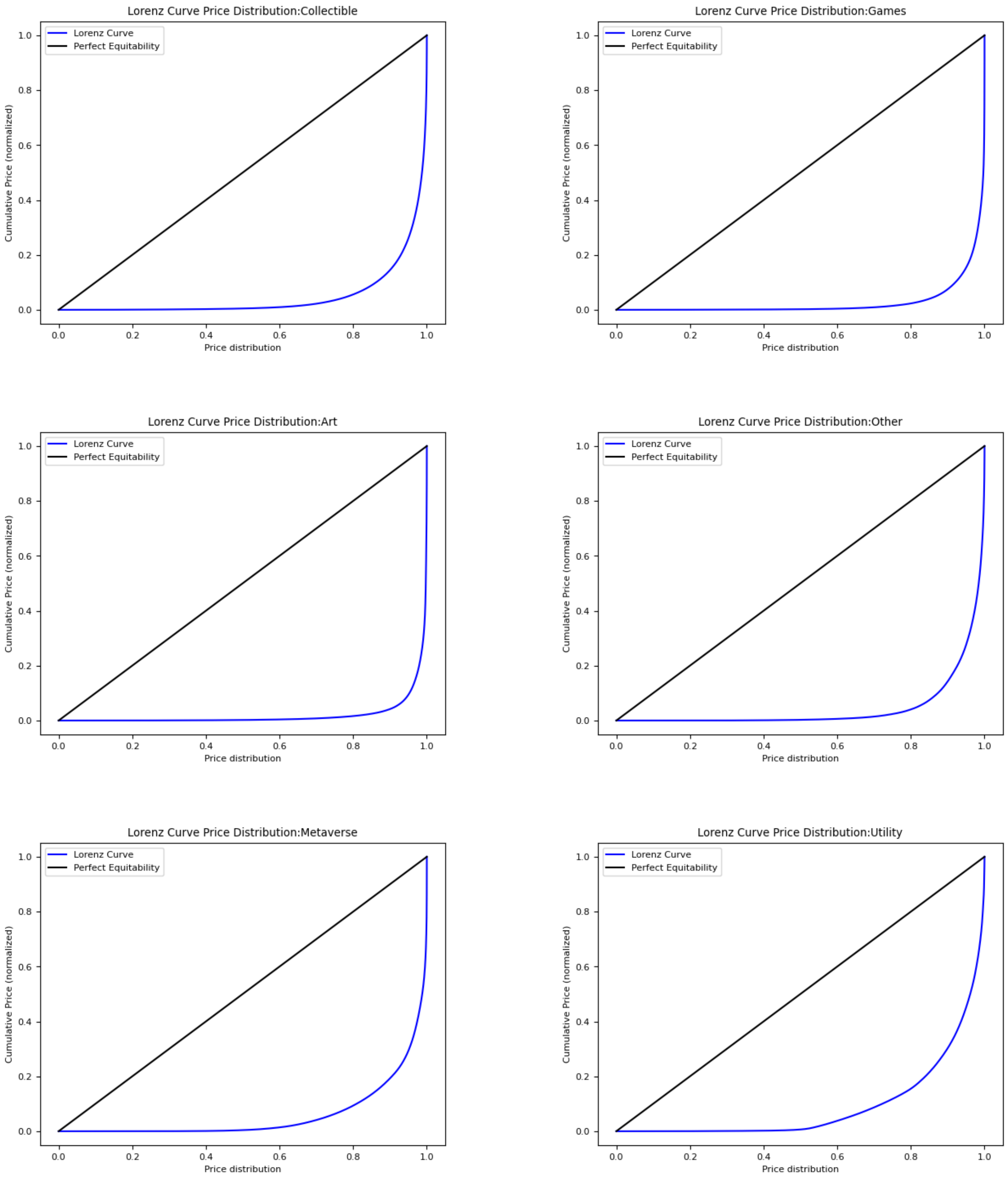


Figure 4.9: Lorenz Curve (By categories)

To compute the Gini Index we calculated the area between the Perfect Equitibility and Lorenz Curve. As mentioned in Chapter 3.1, the Gini Index is a measure of income inequality, and ranges between 0 and 1. A value of 0 indicates perfect equality, where everyone has the same income, while a value of 1 indicates perfect inequality, where a group of NFTs has all the wealth and everyone else has none. From the Table 4.8, we can see that the overall Gini Index value for the NFT dataset is 0.481630, which suggests a relatively high level of inequality in terms of prices across the different NFT categories. The highest Gini Index values are observed for the Art and Games categories, with values of 0.480957 and 0.473247, respectively. The lowest Gini Index values are observed for the Utility and Metaverse categories, with values of 0.408745 and 0.442626, respectively. This suggests that these categories have a relatively more equal distribution of prices compared to the other categories.

Table 4.8: Gini Index by Category

Category	**Gini Index**
All categories	0.481630
Art	0.480957
Games	0.473247
Other	0.455134
Collectible	0.453071
Metaverse	0.442626
Utility	0.408745

4.4 Time-series Analysis and Moving Average

In this section, we examine the changes in average price and volume of NFT transactions over time using time-series analysis and moving average techniques. We first present the general trends in average price and volume change over time for all categories, followed by a breakdown of these trends by category. The figures in this section are computed by grouping the data by the dates of the sale, computing the daily average price in USD and the number of transactions per day (volume), and normalizing the computed numbers between 0 and 1. The x-axis represent every day from June 23, 2017, to April 27, 2021 totalling up to 6,293 days, and the y-axis represents the change of these two dimensions over time.

The Fig 4.10 illustrates that both the average price and volume of NFTs have been steadily increasing since 2017, with a peak in early 2021. The increase in both metrics suggests a growing interest in NFTs among investors and collectors. Additionally, the chart shows that there are some periods of fluctuation, indicating that the NFT market is not immune to volatility. The figure also provides insights into the relationship between average price and volume, as the two metrics seem to follow similar patterns with the trend of

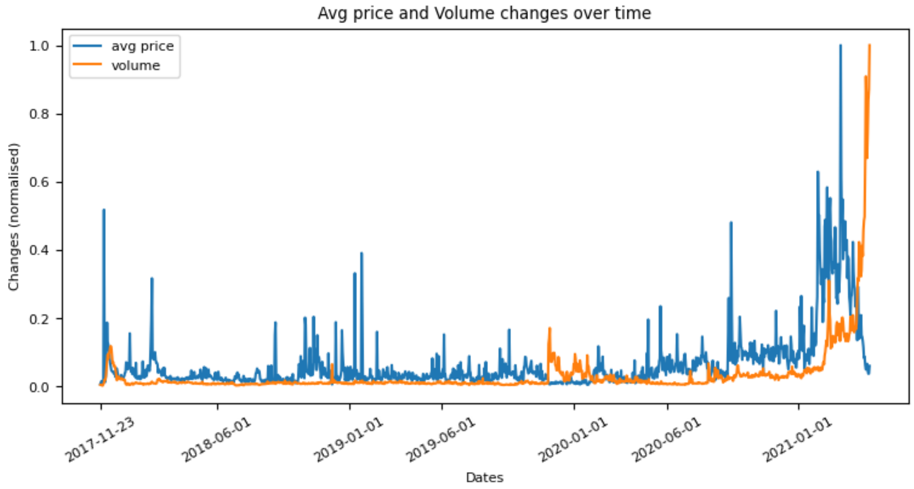


Figure 4.10: Changes in Average Price and Volume over time

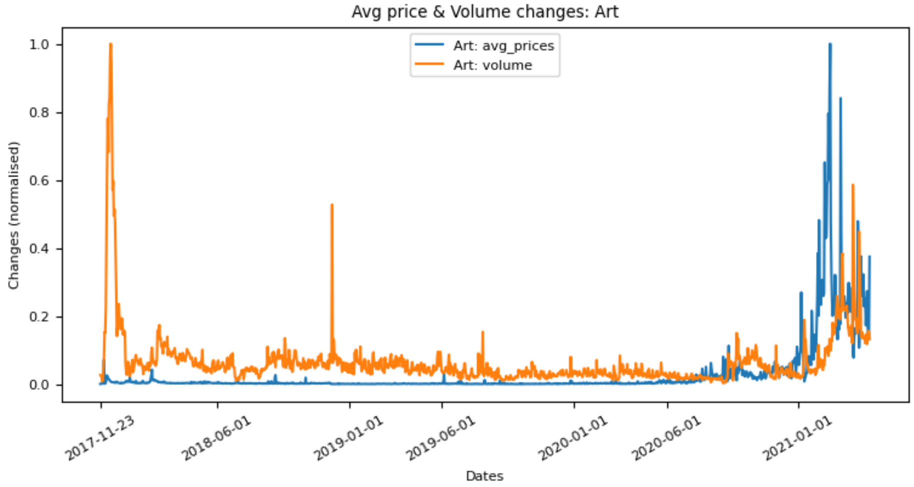


Figure 4.11: Art: Changes in Average Price and Volume over time

volume following the growth rates of the price in 2021. These trends are further analyzed by category in the following sections.

Analyzing the trends by category, we can notice that the earliest NFT product that appeared in the market is Art. However, compared to the overall trend on the NFT market, it seems that NFTs sold as Art reached their peak of popularity in late 2017, late 2018, and early 2021 respectively, in terms of the volume. On the other hand, the changes in price over time had its peak post 2021.

The next category of NFT that appeared on the market is Collectible that has an opposite trend compared to Art. The two peaks of the daily average prices appear in the first half of 2018, after that the trend remains low and steady. However, the popularity of the category arises in 2021, similar to the overall trend of all categories (Fig 4.10).

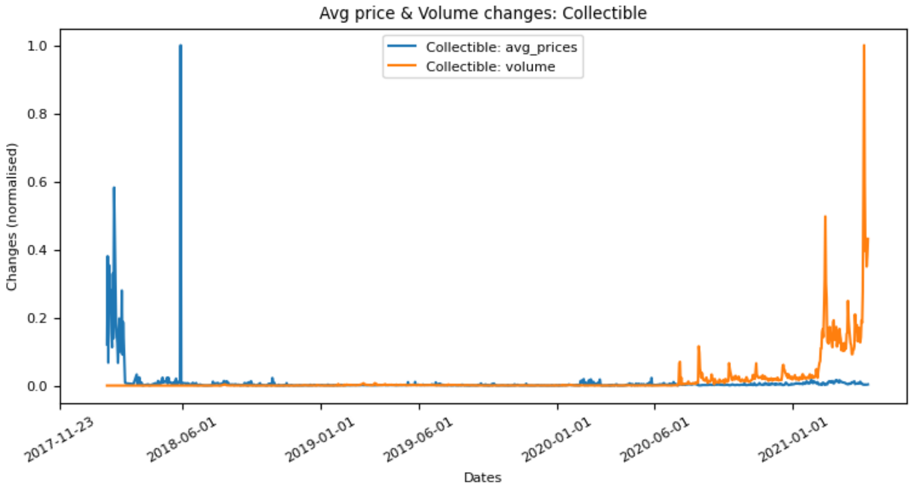


Figure 4.12: Collectible: Changes in Average Price and Volume over time

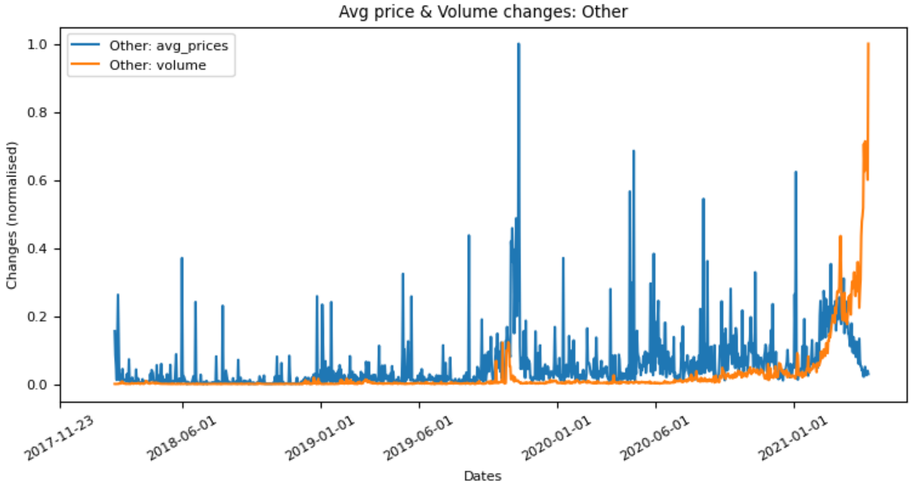


Figure 4.13: Other: Changes in Average Price and Volume over time

The following category that appeared on the market is Other. Similar to Collectible, the trend of the volume for the NFTs of this category goes up after 2021 with a slight popularity rate in late 2019. However, in terms of the daily average price changes, the fluctuations can be observed throughout the whole period with a peak in late 2019.

Another category that appeared on the NFT market is Games. Similar to Other categories, but different from the overall trend of all the categories, the daily average price of the Games NFTs fluctuates throughout the period reaching a few peaks every 5-7 months. Also, similarly to the previous figure, the trend of the volume for these NFTs skyrocketed in the beginning of 2021 with some disturbance in the late 2019.

The next category that appeared on the market in the second half of 2018 is Metaverse. In contrast to Collectible, Other, and Games categories, the popularity of this category hits

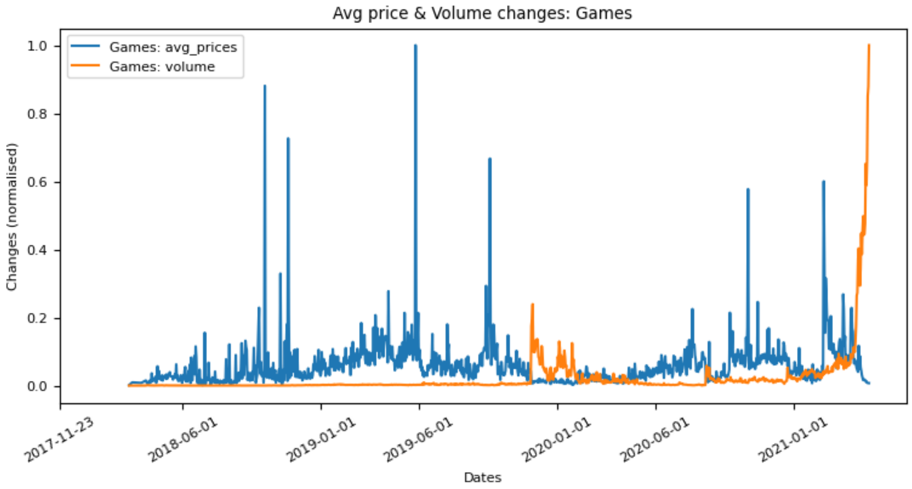


Figure 4.14: Games: Changes in Average Price and Volume over time

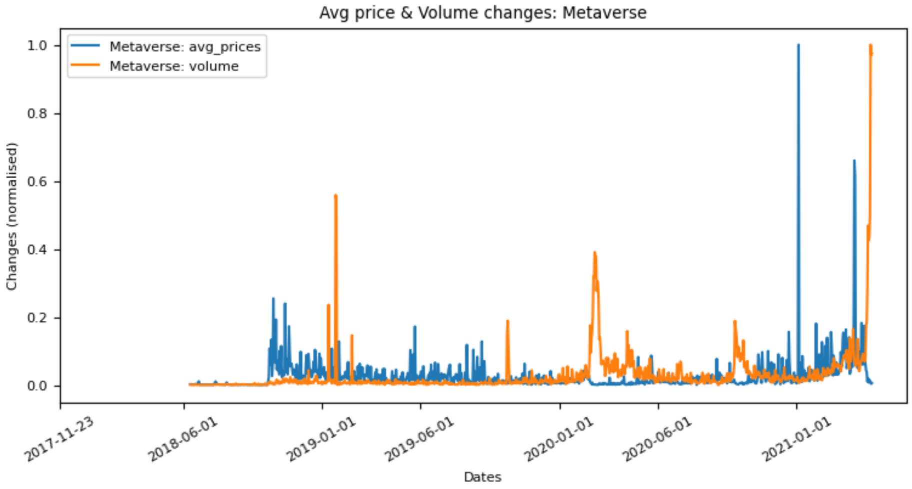


Figure 4.15: Metaverse: Changes in Average Price and Volume over time

the extrema points in early 2019, early 2020, and then early 2021 (once every year). However the daily average price changes fluctuates the whole period reaching its two consecutive peaks after 2021.

The last, but not least category that appeared on the market is Utility. From the figure below we can see that even though the daily average price distribution is fairly static throughout the time, it reaches its peak in the late 2020. However, the popularity of this category reached the highest three points in late 2019, second half of 2020, and past 2021.

The time-series analysis reveals some interesting findings about the changes in average price and volume of NFT transactions. As shown in Fig 4.10, the average price and volume of NFTs have been steadily increasing since 2017, with a peak in early 2021. The overall trend suggests that the NFT market has experienced a significant growth in popularity

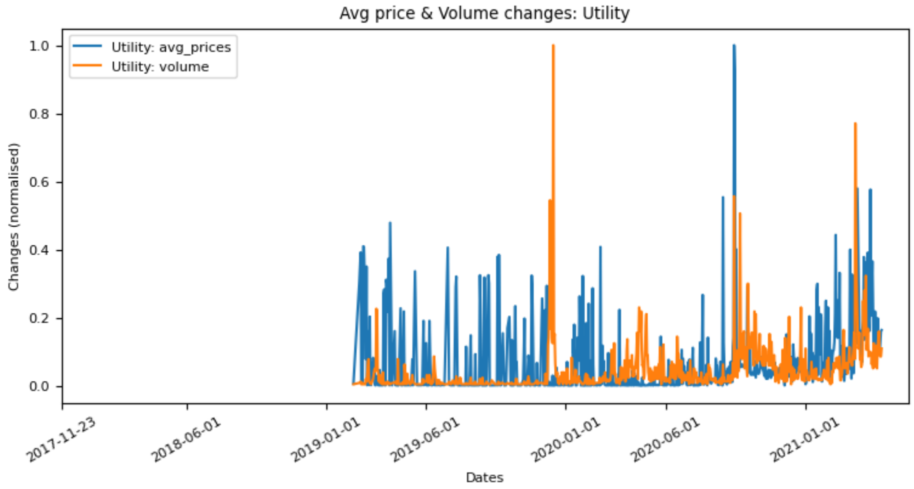


Figure 4.16: Utility: Changes in Average Price and Volume over time

among investors and collectors. However, there are periods of fluctuations that suggest the market is not immune to volatility; therefore, we need to consider techniques that would smooth the overall trends.

Breaking down the trends by category provides further insights. The Art category, the earliest NFT product that appeared in the market, reached its peak of popularity in late 2017, late 2018, and early 2021, respectively, in terms of volume. In contrast, the Collectible category had two peaks in daily average prices in the first half of 2018, but its popularity increased in 2021. The Other category had a slight popularity rate in late 2019 and saw an increase in volume after 2021, while Games had a fluctuating daily average price throughout the period and a surge in volume in early 2021. The Metaverse category experienced its popularity peaks once every year, in early 2019, early 2020, and then early 2021, while the Utility category reached its highest points in late 2019, the second half of 2020, and 2021, with a fairly static daily average price distribution throughout the time. These findings provide a detailed picture of the changes in NFT market trends by category over time, which can be useful for investors and collectors seeking to make informed decisions.

As was discussed in Chapter 3.2, moving averages are commonly used in time-series analysis to identify trends and patterns in data. In the context of the NFT market, moving averages can help us smooth out short-term fluctuations and focus on longer-term trends. By calculating the average price over a specific period, we can obtain a more accurate representation of the market's direction and momentum. Moving averages can be particularly useful when analyzing NFTs, which can experience significant volatility in price and volume. In this section, we apply moving average techniques to the NFT data to gain deeper insights into the trends and patterns in the market and compare them to the previously analyzed figures. The rolling window size was chosen at 7 days.

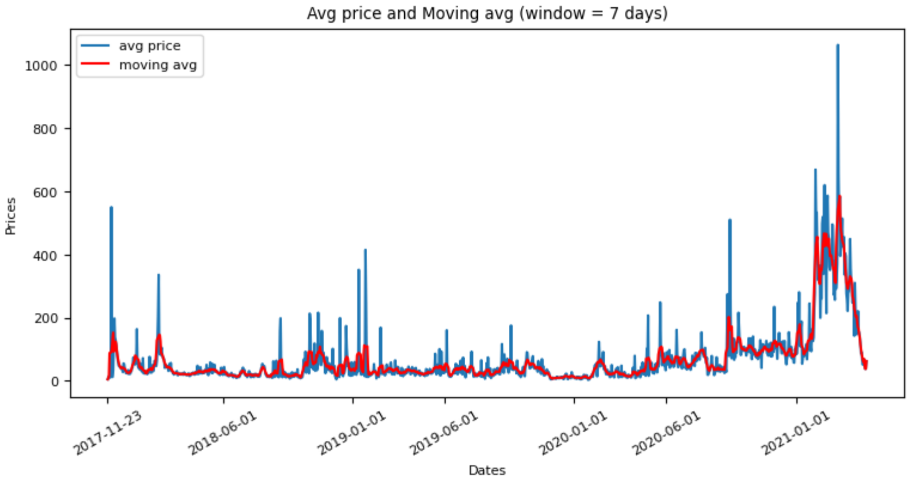


Figure 4.17: Average Price versus Moving Average (7 days)

The overall trend of the average daily pricing for all categories is as follows (see Fig 4.17):

Further, by applying the same technique to every category, we can observe that the trends of the moving average for Art and Collectible are opposite to each other: while the category of Art has an overall increasing trend, the Collectible has a decreasing trend. Similar to Art, the moving average of Metaverse also goes up, but with less significant effect. However, the categories of Other, Games, and Utility still demonstrate price fluctuations throughout the whole period.

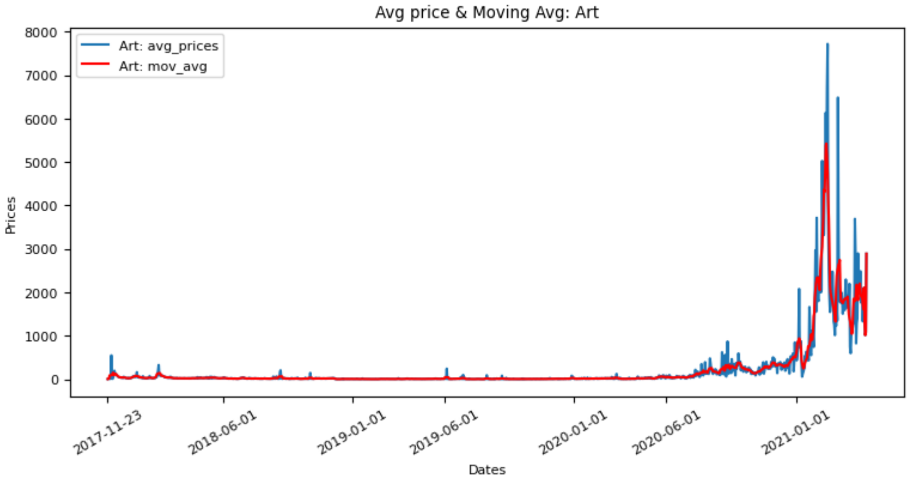


Figure 4.18: Art: Average Price versus Moving Average (7 days)

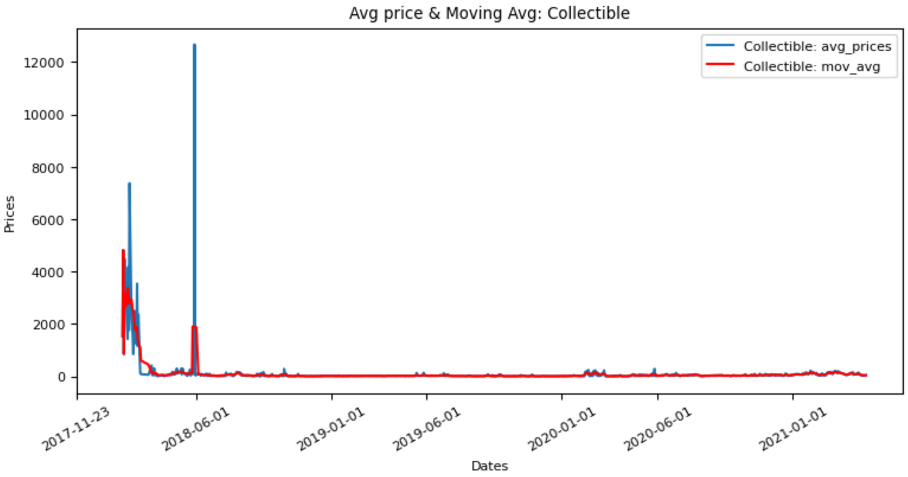


Figure 4.19: Collectible: Average Price versus Moving Average (7 days)

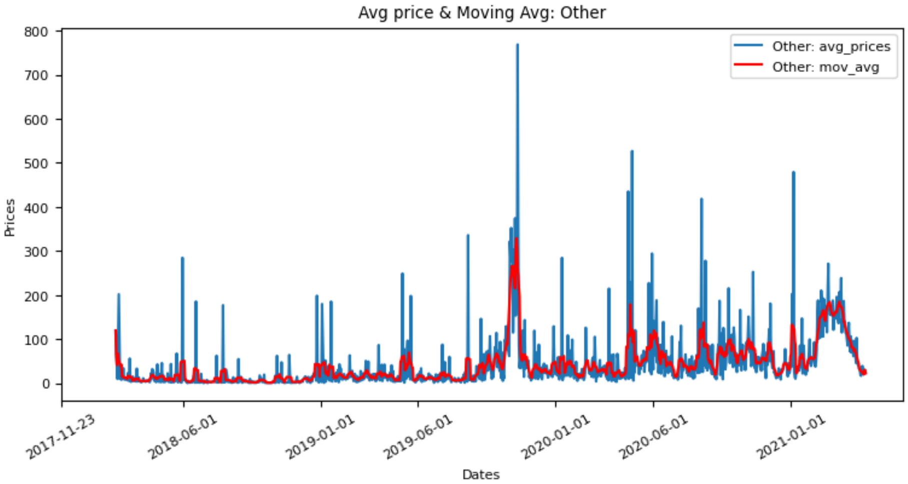


Figure 4.20: Other: Average Price versus Moving Average (7 days)

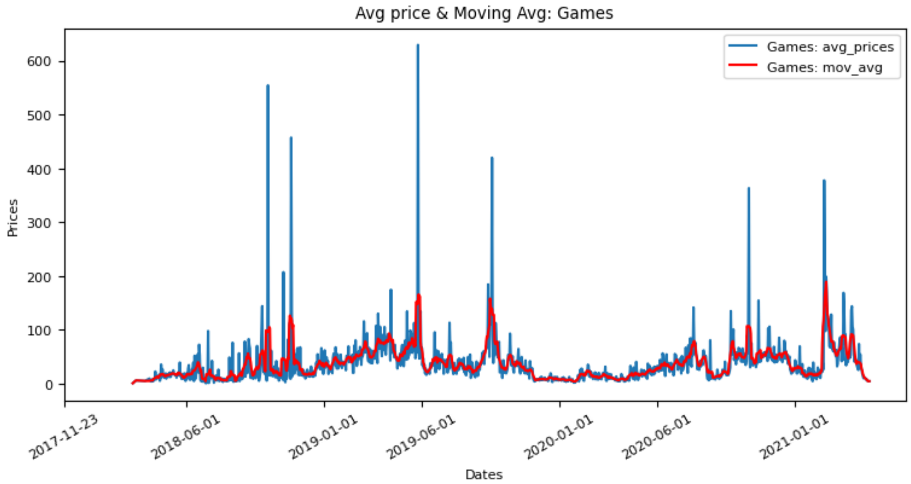


Figure 4.21: Games: Average Price versus Moving Average (7 days)

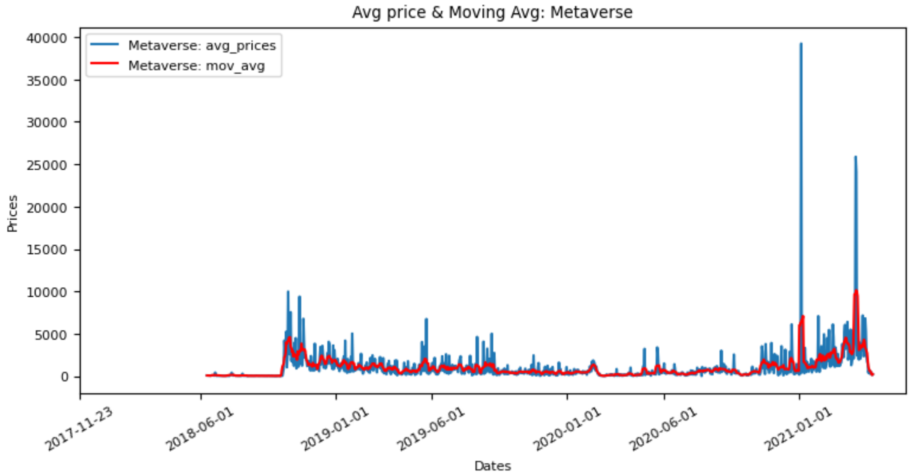


Figure 4.22: Metaverse: Average Price versus Moving Average (7 days)

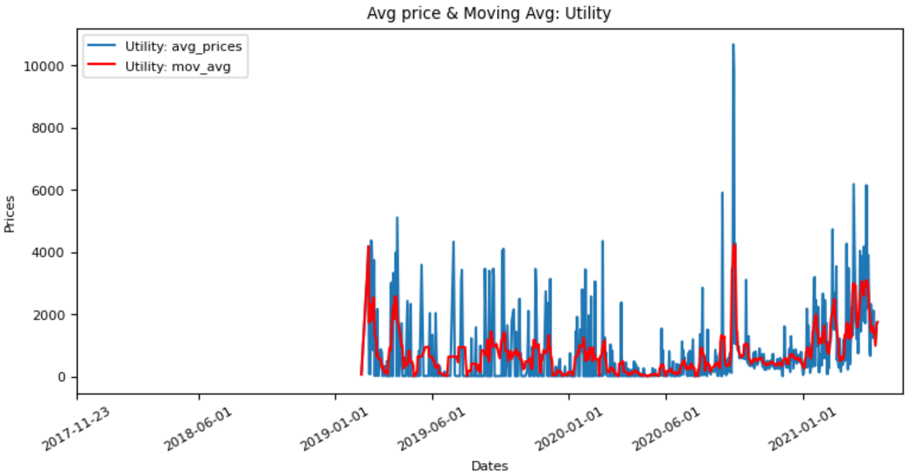


Figure 4.23: Utility: Average Price versus Moving Average (7 days)

4.5 Correlations

In this section, we focus on analyzing the correlation between the six categories of NFTs on the market. To conduct this analysis, we first filtered our dataset to include only the dates when all six categories were being sold on the market, which resulted in a total of 794 days between February 11, 2019, and April 27, 2021. Next, we computed the daily average price in USD and volume for each category.

We then used the Pearson correlation coefficient to measure the strength of the linear relationship between each pair of categories. The Pearson correlation coefficient ranges between -1 and 1, with a value of 1 indicating a perfect positive linear correlation, a value of -1 indicating a perfect negative linear correlation, and a value of 0 indicating no correlation. Finally, we visualized the correlation matrix using a heatmap to identify any significant relationships between the categories. The following paragraphs provide a detailed analysis of our findings.

Looking at Table 4.9 and Fig 4.24, we can see that Collectibles and Art prices have the highest correlation coefficient of 0.41, followed closely by Other, Art and Utility at 0.29. This suggests that there is a weak positive correlation between these categories, meaning that they tend to move in relatively similar directions. On the other hand, the rest of the categories do not have any significant correlations in terms of the price change over time.

From the correlation matrix (Table 4.10 and Fig 4.25), we can see that the volume of transactions in Collectible, Games, and Other categories have a relatively high correlation with each other, with correlation coefficients ranging from 0.71 to 0.86. Meanwhile, the volume of transactions in Art and Metaverse categories have a relatively low correlation with the other categories, with correlation coefficients ranging from 0.24 to 0.65. The volume

Table 4.9: Correlation Matrix of Prices

Prices	Art	Collectible	Games	Metaverse	Other	Utility
Art	1.000000	0.414549	0.230264	0.264809	0.290189	0.285110
Collectible	0.414549	1.000000	0.122678	0.241852	0.200546	0.192540
Games	0.230264	0.122678	1.000000	0.146260	0.110030	0.181031
Metaverse	0.264809	0.241852	0.146260	1.000000	0.123646	0.163100
Other	0.290189	0.200546	0.110030	0.123646	1.000000	0.069505
Utility	0.285110	0.192540	0.181031	0.163100	0.069505	1.000000

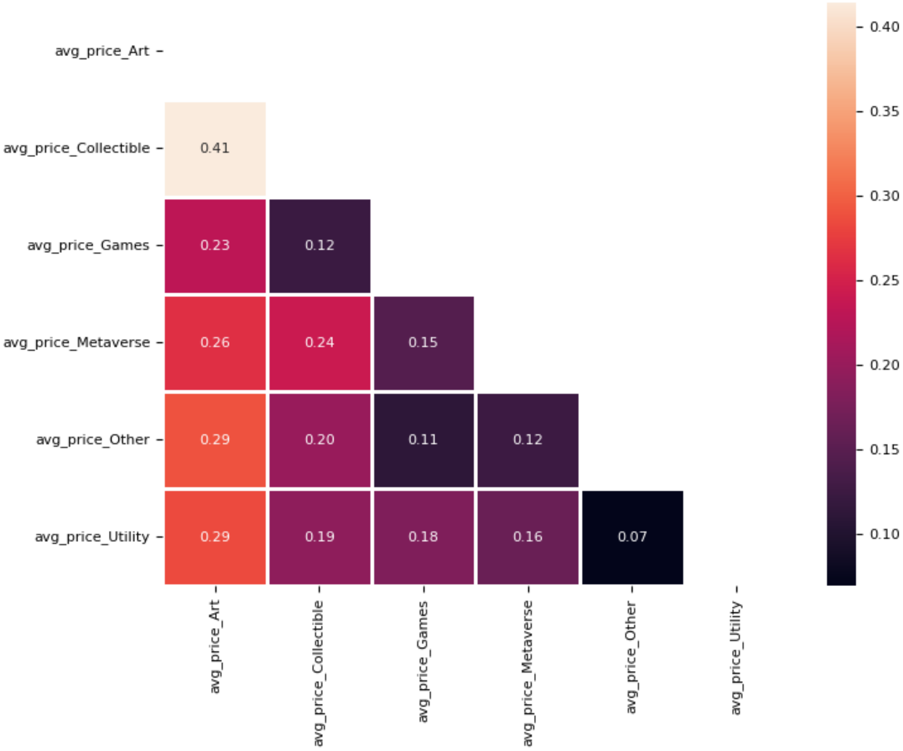


Figure 4.24: Correlation Matrix between different categories

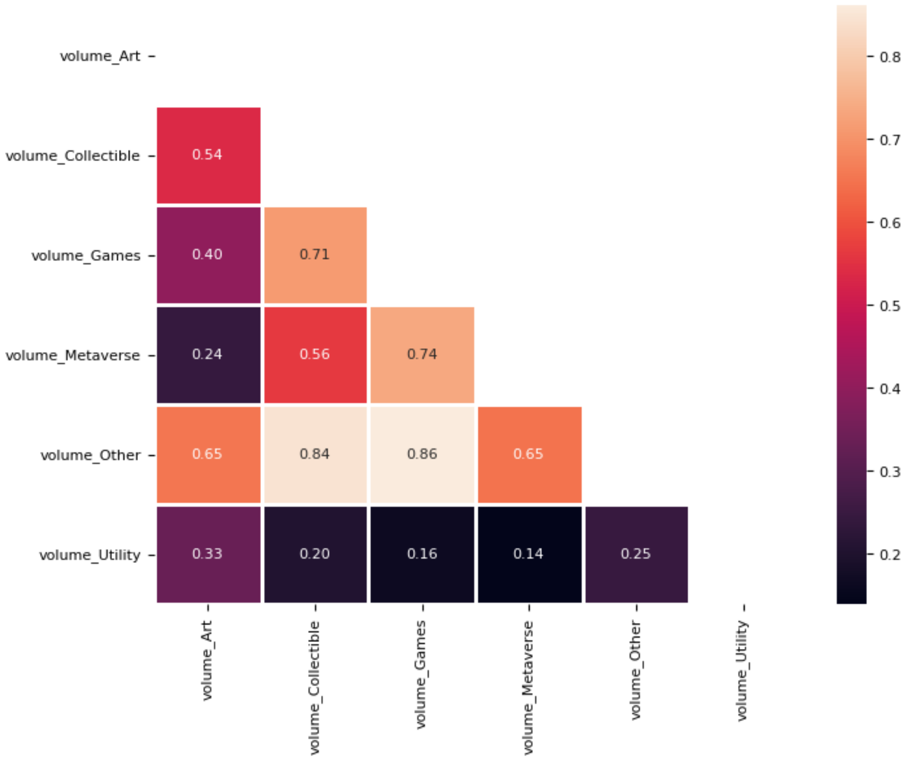


Figure 4.25: Correlation Matrix between different categories

of transactions in the Utility category has a very low correlation with all other categories, with correlation coefficients ranging from 0.14 to 0.33.

Table 4.10: Correlation Matrix - Volume

Category	**Art**	**Collectible**	**Games**	**Metaverse**	**Other**	**Utility**
Art	1.000000	0.536062	0.400288	0.240827	0.652706	0.331847
Collectible	0.536062	1.000000	0.711311	0.564013	0.843593	0.202956
Games	0.400288	0.711311	1.000000	0.736215	0.860377	0.163906
Metaverse	0.240827	0.564013	0.736215	1.000000	0.647091	0.138643
Other	0.652706	0.843593	0.860377	0.647091	1.000000	0.245240
Utility	0.331847	0.202956	0.163906	0.138643	0.245240	1.000000

Chapter 5

Results: Hierarchy of Clusters

As discussed in Chapter 3.4, the purpose of constructing the hierarchy of clusters was to organize various types of NFTs into distinct clusters and explore the most prominent ones using K-means clustering. Given the wide range of NFT prices, it was hypothesized that certain clusters may be more extensive than others and could be analyzed in more detail by removing the smaller clusters from the dataset. This approach enabled us to focus on the most significant trends and insights, while still considering the broader patterns across all NFTs.

5.1 Feature selection

After conducting a thorough examination of the NFT dataset, we extracted five main features to use in our analysis: *local_avg_price*, *global_avg_price*, *counts_norm*, *local_std*, and *global_std*. To calculate *local_avg_price*, we normalized the sale price of each NFT within its own sale history between 0 and 1, and then found the average of these normalized values. For *local_std*, we calculated the standard deviation of the sale prices within the same sale history. To obtain *global_avg_price* and *global_std*, we first calculated the average and standard deviation of the sale prices for each NFT across all its sales and then scaled these values between 0 and 1 using the minimum and maximum values across all NFTs. Finally, to calculate *counts_norm*, we determined the number of times each NFT was sold and normalized this value between 0 and 1 using the minimum and maximum values across all NFTs.

When calculating each feature, we assume that n is a number of sales per NFT, where P_i is the price of the i-th sale, p_{max} is the maximum price in the sale history, μ is local average price, N is total number of NFTs, P_j is the average price of the j-th NFT across all its sales, P_{max} is the maximum price across all NFTs, C is the number of times an NFT was sold, C_{min} is the minimum number of times any NFT was sold, and C_{max} is the maximum number of times any NFT was sold:

$$local\_avg\_price(\mu) = \frac{1}{n}\sum\frac{P_i}{p_{max}} \tag{5.1}$$

$$local\_std = \sqrt{\frac{1}{n-1}\sum\frac{P_i}{p_{max}-\mu}^{2}} \tag{5.2}$$

$$global\_avg\_price(\nu) = \frac{1}{N}\sum(\frac{P_j}{P_{max}}) \tag{5.3}$$

$$global\_std = \sqrt{\frac{1}{N-1}\sum\frac{P_j}{P_{max}-\nu}^{2}} \tag{5.4}$$

$$counts\_norm = \frac{C - C_{min}}{C_{max} - C_{min}} \tag{5.5}$$

The above formulas were used to calculate features per each NFT, however, for the NFTs that were only sold once (vast majority) these features cannot be calculated. Therefore, we initially clusterized the NFTs into 2 clusters at the Root Level: sold once and sold many times.

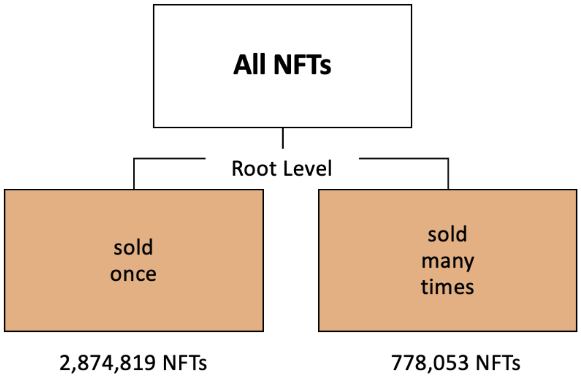


Figure 5.1: Root Level clustering

5.2 Process of Clustering

As can be observed from Fig 5.2, the process of clustering starts with plotting inertia values for k between 2 and 11, to apply the 'Elbow Method' discussed in Chapter 3.4. Then, we choose the optimal number of k (number of clusters), and apply the K-means algorithm based on the optimal k. After plotting and analyzing the results of an initial clustering, we choose larger clusters (those that constitute over 40% of all NFTs) and repeat the process until all the clusters look balanced.

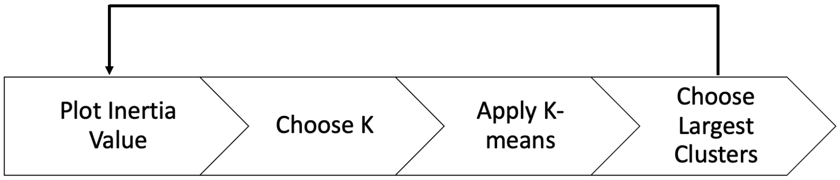


Figure 5.2: Process of clustering

5.3 Initial Results and Triangular-shaped Clusters

Initially, after cleaning the dataset from missing values and other data inconsistencies, we were able to extract features of 778,053 unique NFTs that were sold at least twice. We plotted (see Fig 5.3) K-means values against Inertia Value (SSE measure) to identify the optimal value of k implying the 'Elbow Method', and identified k=3 as the optimal number of clusters.

As our next step, we applied K-Means algorithm with k = 3, and were able to identify 3 main clusters (see Fig 5.4) with the largest cluster 2 (in blue) constituting over 76% of all NFTs sold at least twice (see Table 5.1).

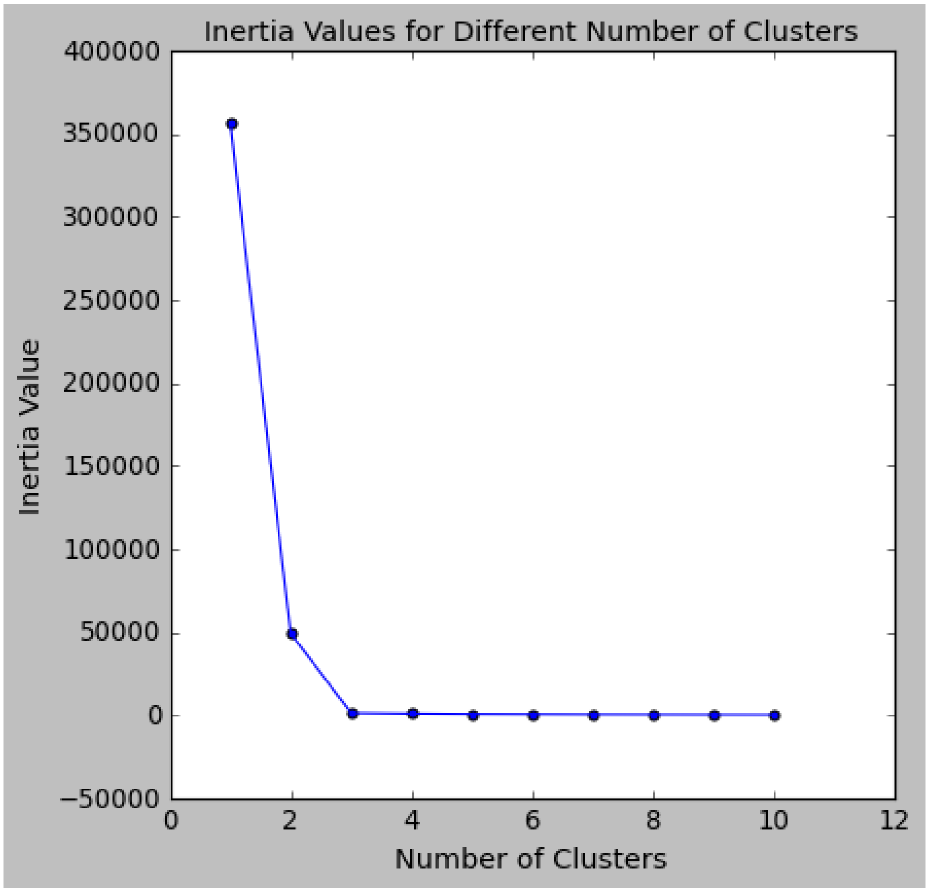


Figure 5.3: Inertia Values versus Number of Clusters

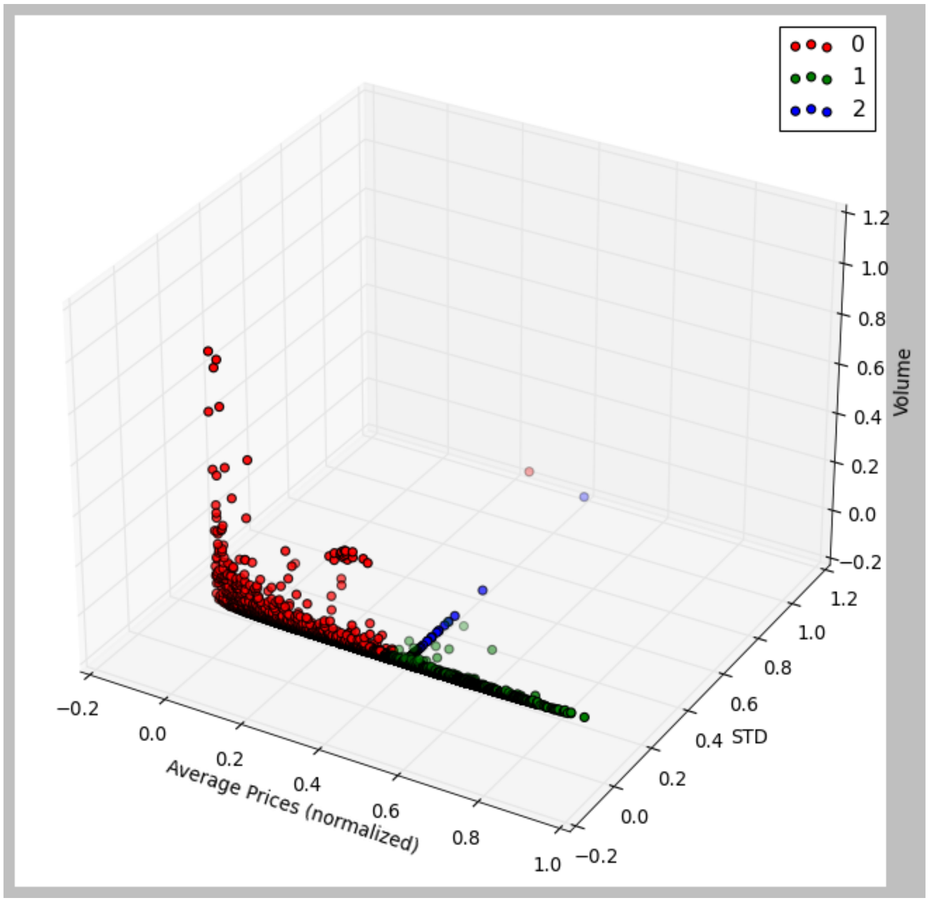


Figure 5.4: Initial Cluster Visualization

Table 5.1: Cluster Analysis Initial Results

Cluster	**Avg Local Price**	**Avg Global Price**	**Counts**	**Local STD**	**Global STD**	**Size**
0	0.357	1.76×10^{-5}	1.41×10^{-4}	0.5131	2.16×10^{-5}	85,670
1	0.5526	1.48×10^{-5}	9.10×10^{-4}	0.5264	8.61×10^{-5}	100,042
2	0.5	0.77×10^{-5}	0	0.7071	5.51×10^{-5}	592,341

Following the procedure described in Section 5.2, we focused on cluster 2, and here are some of our observations:

- The features calculated for local values (local_avg_price, local_std) as well as counts_norm are uniform for all the NFTs in cluster 2. Therefore, we only used the remaining two features (global_avg_price, global_std) for visualizations.
- Analyzing the inner clusters further, we noticed that all subsequent clusters are triangular-shaped (see Fig 5.5), indicating a relationship between the two variables with a line function:

$$\text{Global STD} = 1.413 \times \text{Global Avg Price} + c$$

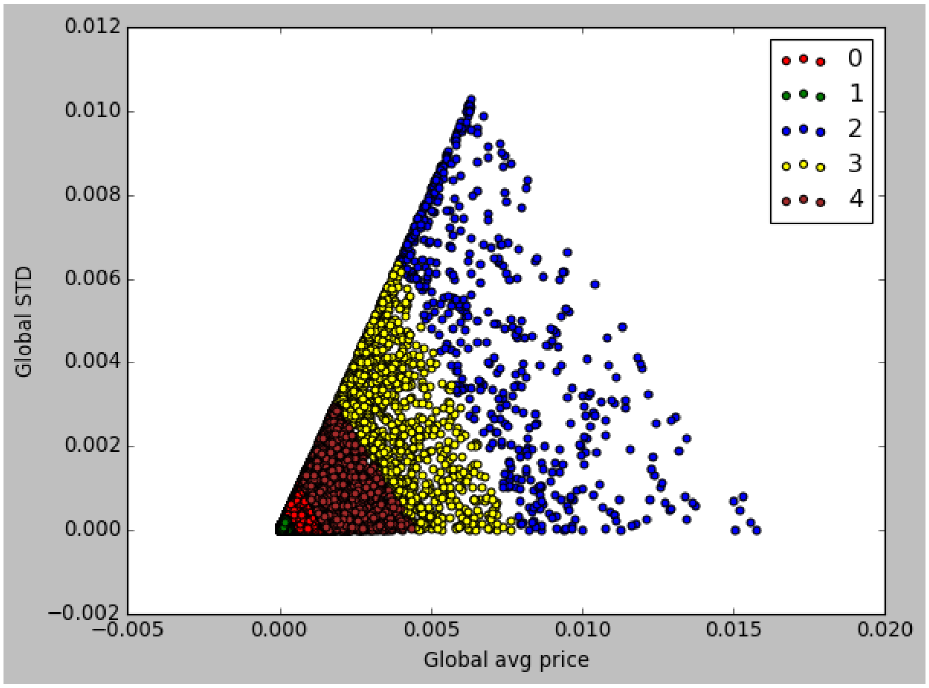


Figure 5.5: Cluster 2 Visualization (2D)

These results looked interesting; therefore, we investigated further and found a trivial explanation to our observations. When an NFT is sold twice, it only has two prices P_1 and P_2, where:

$$avg\_price(\mu) = \frac{1}{2}(P_1 + P_2) \tag{5.6}$$

$$std = \sqrt{\frac{(P_1 - \mu)^2 + (P_2 - \mu)^2}{2}} \tag{5.7}$$

if we define distance from each price P to the mean value as d, then:

$$(P_1 - \mu) = (P_2 - \mu) = d \tag{5.8}$$

$$std = \sqrt{\frac{(d)^2 + (d)^2}{2}} \tag{5.9}$$

therefore, std and μ have a linear relationship, which explains the triangular-shaped clusters.

After analyzing these results, we decided to group our clusters into two initial clusters, and 2 subclusters at the Level 1: sold once (one price available), sold at least two times, sold twice (two prices available), and sold many times (three or more prices available).

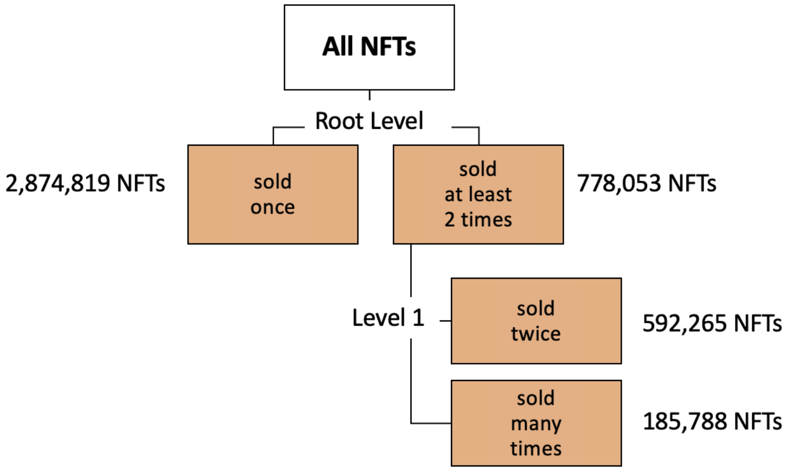


Figure 5.6: Hierarchy of Clusters at Level 1

5.4 Final Results and the Hierarchy of Clusters

In this section, we will focus on building the hierarchy of clusters after excluding the larger clusters of the NFTs that were sold once and sold twice. By calculating the Inertia Values of the remaining 185,788 NFTs in the range of k between 2 and 11, we found our optimal k to be 4 (see Fig 5.7). Applying K-Means algorithm for four clusters resulted in four subclusters (see Fig 5.8)

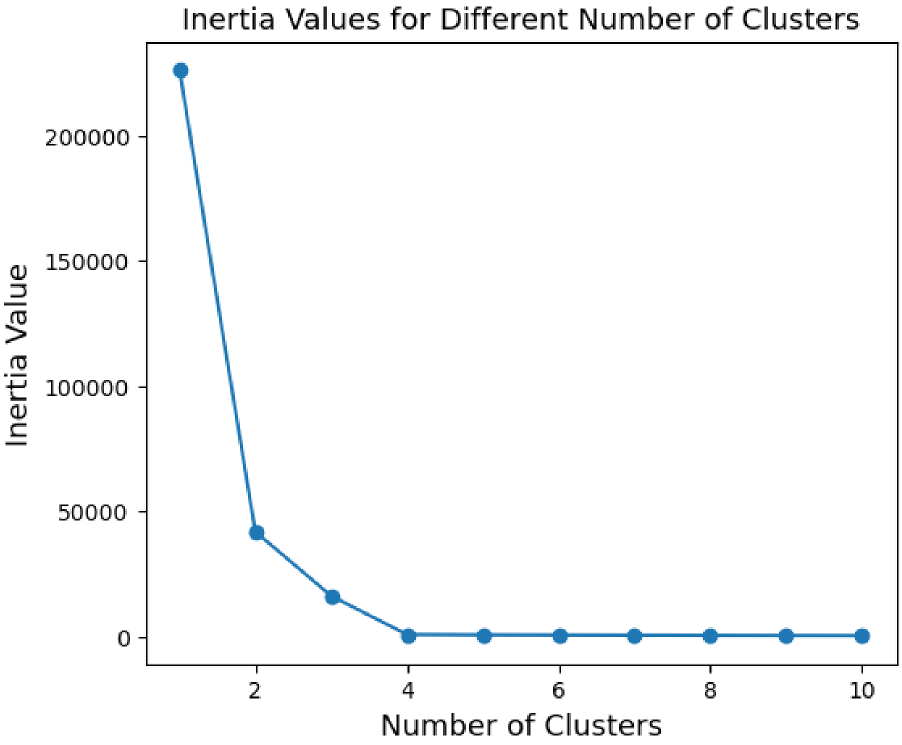


Figure 5.7: Inertia for different values of k

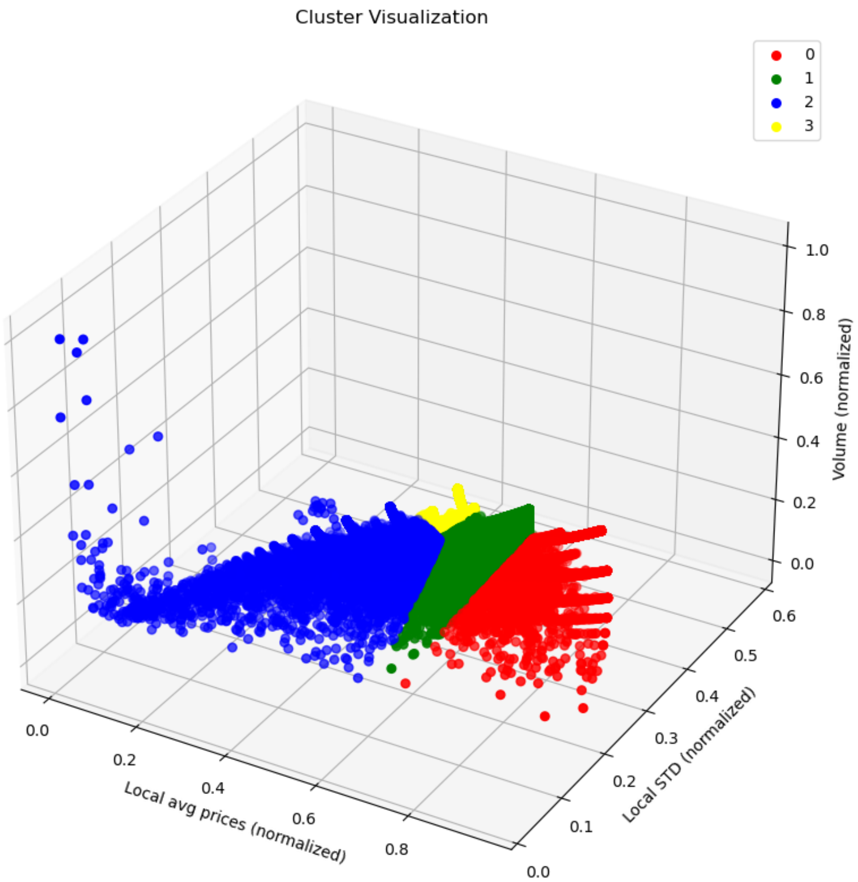


Figure 5.8: Cluster Visualization

Table 5.2: Observations for Different Clusters

Cluster	Avg Local Price	Avg Global Price	Counts	Local STD	Global STD	Size (Percent)
2	0.306	2.81×10^{-4}	0.423	3.89×10^{-4}	3.16×10^{-4}	21,519 (11.59%)
0	0.633	1.69×10^{-4}	0.531	0.99×10^{-4}	1.95×10^{-4}	37,642 (20.27%)
3	0.357	1.40×10^{-4}	0.557	1.72×10^{-4}	0.54×10^{-4}	50,722 (27.31%)
1	0.492	1.36×10^{-4}	0.518	0.82×10^{-4}	4.41×10^{-4}	75,829 (40.83%)

Analyzing Fig 5.8 and Table 5.2, we can characterize each cluster as follows:

- Cluster-2: even though the local average price of these NFTs is the smallest, comparing it to the rest of the NFTs in other clusters, their global average prices are much higher. In terms of the price distribution of this particular cluster, both local and global STD is higher than in any other class. We can generalize this group as NFTs that are very expensive, have lower rate of resale, and with a large range of prices, but the median of the prices internally is closer to its minimum price.

- Cluster-0: has the largest average price rate locally, meaning that their average price internally was closer to a maximum price, but has a lower price globally compared to cluster-2. At the same time, this group of NFTs along with cluster-2 and cluster-1, are sold more often than cluster-2. Cluster-0 also has a low STD locally, meaning that the prices for which these NFTs were sold changed gradually and didn't differ too much from its original price. Finally, this group of NFTs has a higher rate of global STD compared to local STD. This means that the NFTs within this cluster less significantly than in cluster-2.

- Cluster-3: has very similar characteristics as cluster-0, but with lower local average price, meaning that their average price internally was closer to a minimum price. Moreover, cluster-3 has a higher local STD than global STD, meaning that all the NFTs in this cluster are in a relatively similar price range, while price changes internally are more significant than between different NFTs.

- Cluster-1: has the largest portion of all NFTs and constitutes 40.83% of all clusters. It also has the smallest global average price and a relatively small volume rate. This cluster will be analyzed further.

The resulting hierarchy after level 2 can be seen below (Fig 5.9). Now, we will look closer into cluster-1.

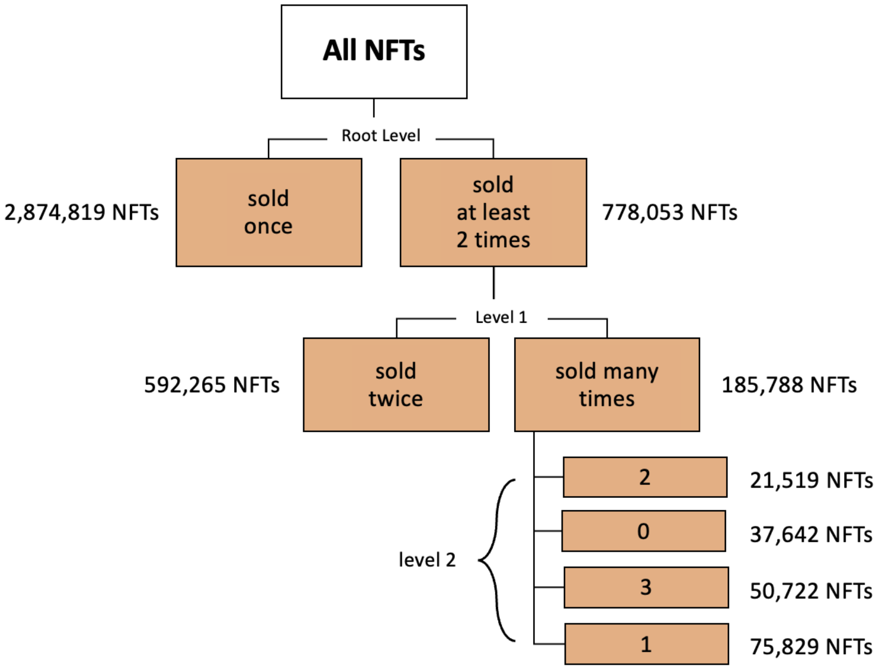


Figure 5.9: Cluster Visualization

At the final stage, we applied the same technique to cluster-1. We first chose a number of optimal clusters to be four based on Fig 5.10. Then, we used K-means clustering to identify subclusters (see Fig 5.11 and Table 5.3) of the cluster-1

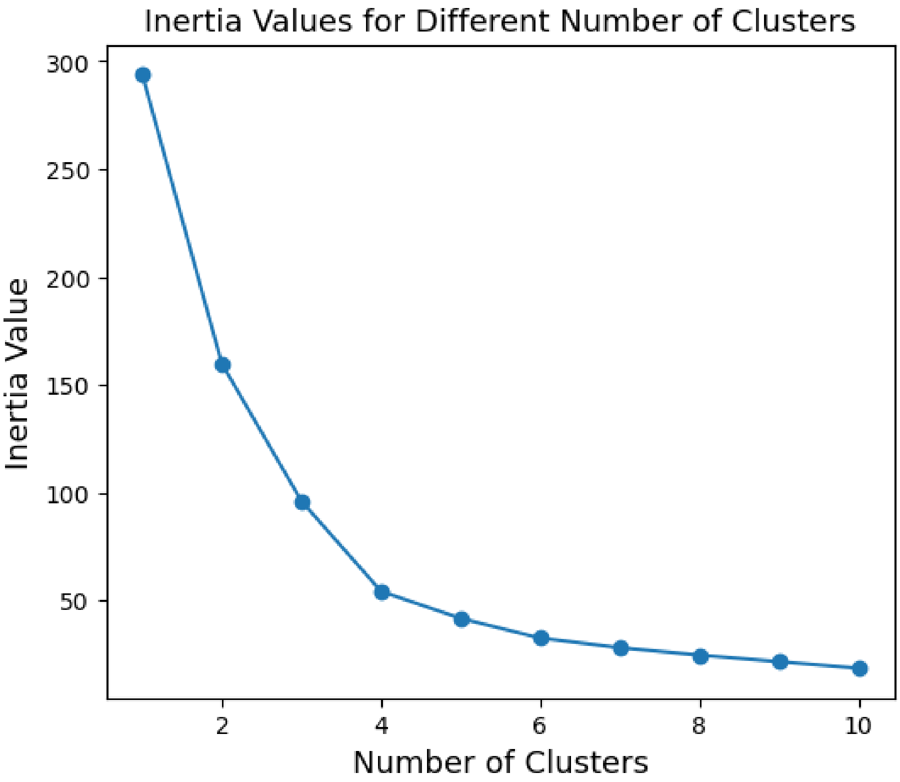


Figure 5.10: Inertia for different values of k

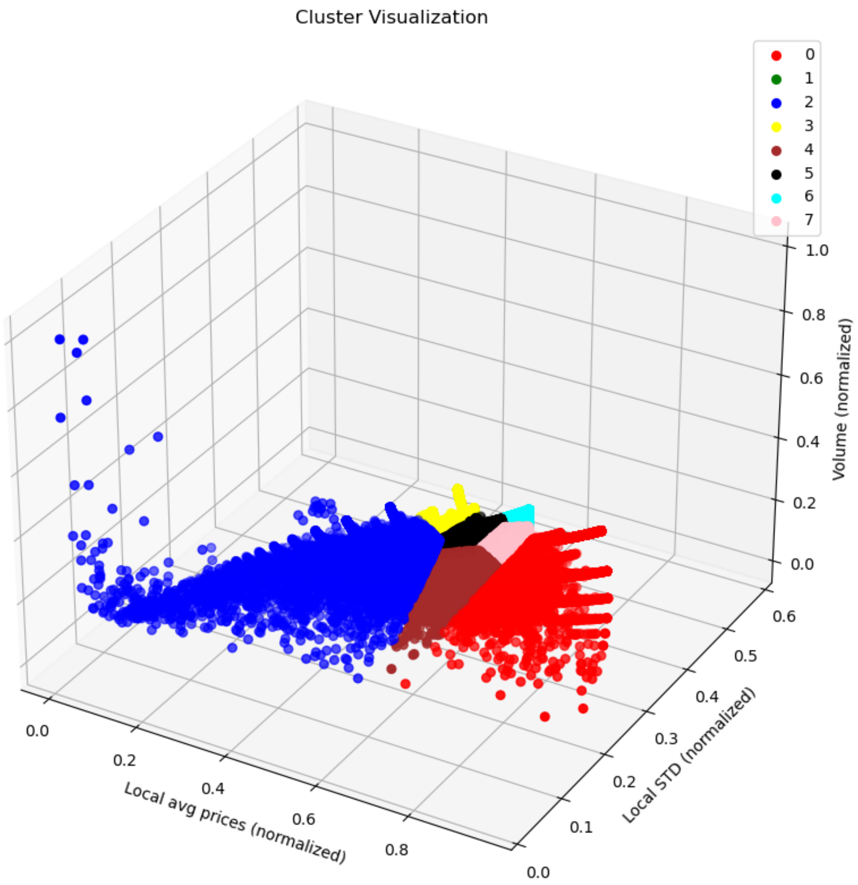


Figure 5.11: Cluster Visualization

Table 5.3: Cluster Analysis Results

Cluster	Avg Local Price	Avg Global Price	Counts	Local STD	Global STD	Size (Percent)
4	0.496	2.32×10^{-4}	0.417	1.48×10^{-4}	1.06×10^{-3}	8,999 (11.87%)
7	0.526	1.52×10^{-4}	0.501	0.93×10^{-4}	0.18×10^{-3}	17,953 (23.68%)
5	0.452	1.47×10^{-4}	0.503	1.01×10^{-4}	0.26×10^{-3}	20,974 (27.66%)
6	0.499	0.86×10^{-4}	0.573	0.41×10^{-4}	0.54×10^{-3}	27,903 (36.79%)

Analyzing Fig 5.11 and Table 5.3, we can characterize each cluster as follows:

- Cluster-4: has a similar pattern with cluster-2 in terms of the global price and volume (counts) metrics, the global average price is large, but the volume is small. This cluster also has the largest values of STD both locally and globally, meaning that price ranges in this group of NFTs are relatively spread.

- Cluster-7: has similar trends with cluster-5 and cluster-6, but its local price measurement is relatively bigger than for other clusters, underlying that the center of the NFTs under cluster-7 are closer to its maximum point.

- Cluster-5: has an opposite trend compared to cluster-7 in terms of the local average price, meaning that its center is closer to a minimum price, even though the global average of the prices for this cluster is smaller than cluster-4 and cluster-7
- Cluster-6: has the smallest value in terms of global average price, but the largest number of transactions.

To summarize the hierarchy of clusters, there is a total of nine classes of NFTs: at the root level and level 1, the classes were separated due to feature extraction peculiarities, while at levels 2 and 3, the separation was realized through K-means hierarchical clustering (see Fig 5.12). From table 5.4 we can also see the original measures of each class in USD, and here are some general observations for each class:

- Class 2 has the highest average price, highest mean STD, and highest mean count, indicating that it has the most valuable and popular NFTs.
- Class 4 has a lower average price but a relatively high mean STD, indicating that it has NFTs with higher price volatility.
- Class 0 has a lower average price and mean STD, but it has a higher mean count compared to other classes, indicating that it has a larger number of lower-priced NFTs.
- Class 7 has a lower average price, mean STD, and mean count than the other classes, indicating that it has a relatively smaller number of lower-priced and lower-volatility NFTs.
- Class 5 has a lower average price and mean STD, but a slightly higher mean count than Class 7, indicating that it has a larger number of lower-priced and lower-volatility NFTs than Class 7.
- Class 3 has a relatively low average price and a high mean STD, indicating that it has NFTs with higher price volatility.
- Class 6 has the lowest average price, lowest mean STD, and lowest mean count, indicating that it has the least valuable and popular NFTs. However, it has the highest STD of prices, which suggests that it may have NFTs with a wider range of prices.

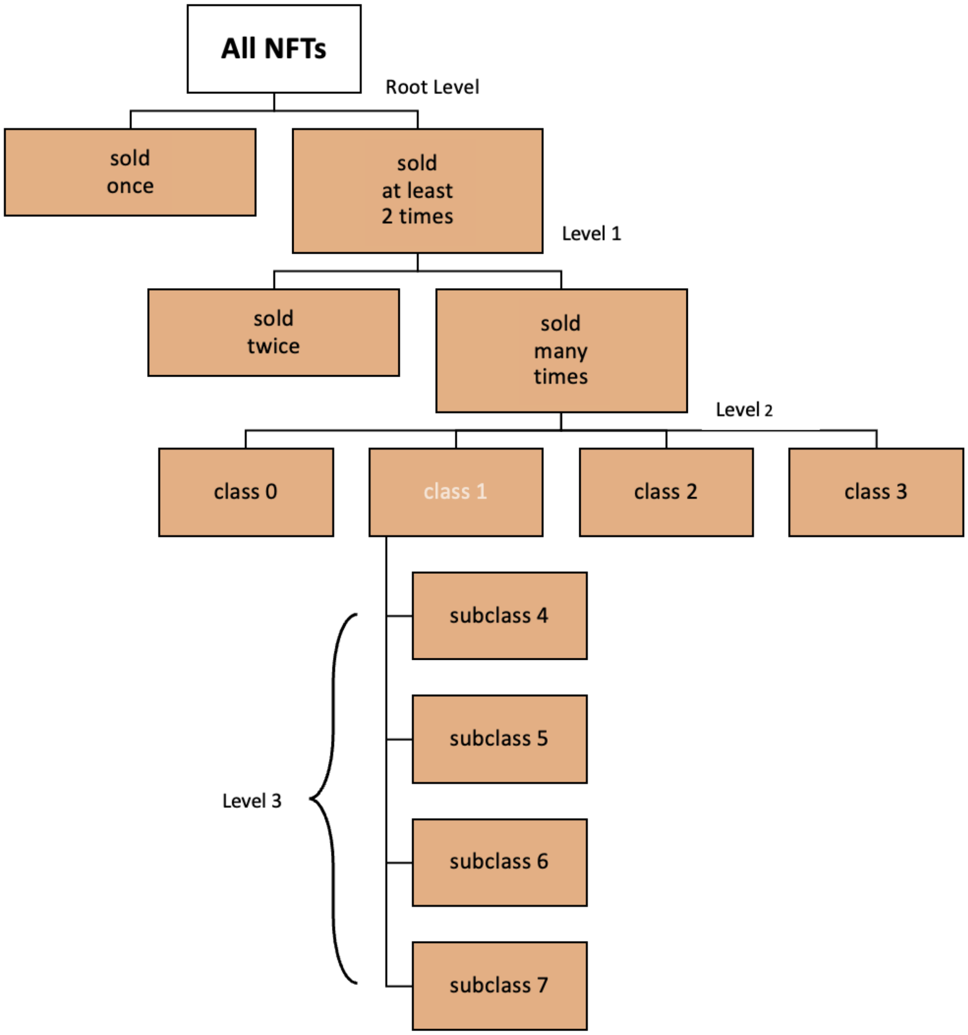


Figure 5.12: Hierarchy of Clusters

Table 5.4: Class Statistics

Classes	Avg Price	Mean STD	Mean Counts	STD of Prices	STD of STD	STD of Counts
2	144.594	174.226	8.820	881.38	1731.40	41.42
4	119.208	66.334	4.956	1507.39	1107.14	1.69
0	87.007	44.298	3.359	664.69	467.95	1.02
7	77.994	41.485	3.343	788.63	512.86	0.69
5	75.629	45.152	3.472	1103.99	484.48	0.77
3	71.886	76.970	3.100	1239.08	2065.67	0.64
6	44.295	18.278	4.000	3139.70	680.29	0.05

Chapter 6

Results: Predicting Prices

In this section, we discuss the process of predicting NFT prices. We first discuss feature selection, followed by the results of the linear regression model, regularized regression, and MLP.

6.1 Feature Selection and Data Preprocessing

During the process of feature selection, we endeavored to extract relevant features for the purpose of predicting the last sale price of non-fungible tokens (NFTs). From an initial pool of over 30 different features, we narrowed down our focus to 11 features that were deemed to be the most significant. Our approach involved a focus on the historical transactional data of NFTs, and specifically, we selected 135,812 NFTs that had been traded at least 5 times.

To preprocess the data, we grouped NFTs and normalized their prices between 0 and 1 within their sale history. We then extracted the last sale price for each NFT as our dependent variable (y value). Additionally, we extracted the prices and dates for the second to last, third to last, fourth to last, and primary sales (see Fig 6.1). In order to capture the variability of the data, we included the average normalized price and standard deviation of these prices as two additional features (see the description of each feature in table 6.1).

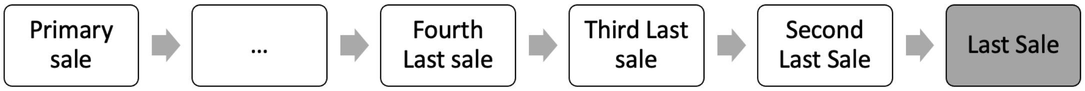


Figure 6.1: Hierarchy of Clusters

Overall, this feature selection process allowed us to narrow down a large pool of potential features to a set of key variables that could be used for the price prediction of NFTs. This served as the foundation for subsequent modeling approaches, including linear regres-

Table 6.1: Feature Descriptions

Feature	Description
global__avg__norm__price	average price of the NFT after normalizing
global__std__norm__price	standard deviation of the NFT prices after normalizing
primary__norm__price	normalized price of the first ever sale
primary__sale__year	year in which the first sale has been made
second__last__norm__price	normalized price of the second to last sale
second__last__sale__year	year in which the second to last sale has been made
third__last__norm__price	normalized price of the third to last sale
third__last__sale__year	year in which the third to last sale has been made
fourth__last__norm__price	normalized price of the fourth to last sale
fourth__last__sale__year	year in which the fourth to last sale has been made
last__sale__year	year in which the last sale is made

sion, regularized regression, and a neural networks model using the Multi-Layer Perceptron (MLP) algorithm. Additionally, we used a 9:1 train/test ratio for the model training and evaluation processes. The models used in this study incorporated 10 cross-fold validation techniques for regression analysis, as well as used validation sets for MLP model evaluation.

6.2 Linear Regression and Regularization

Simple Linear Regression: In this approach, we used a linear regression model to predict the last sale price of an NFT based on a set of selected features. We used a 10-fold cross-validation method to evaluate the performance of our model. The performance was measured using the Root Mean Squared Error (RMSE) metric, which indicates the difference between the predicted and actual values. We also calculated the coefficients of the features in the linear regression model, which indicates how strongly each feature affects the predicted value.

L1 Regularized Regression (Lasso): In this approach, we used Lasso regularization to reduce the complexity of our model and prevent overfitting. We first found an optimal alpha value using a 10-fold cross-validation method. We then used this alpha value to train our Lasso regression model and calculated the RMSE and standard deviation of our predictions. We also calculated the coefficients of the features in the Lasso regression model.

L2 Regularized Regression (Ridge): In this approach, we used Ridge regularization to reduce the complexity of our model and prevent overfitting. Similar to Lasso, we first found an optimal alpha value using a 10-fold cross-validation method. We then used this alpha value to train our Ridge regression model and calculated the RMSE and standard deviation of our predictions. We also calculated the coefficients of the features in the Ridge regression model, which indicates the importance of each feature in the prediction. Table

6.2 demonstrates the coefficient values for each model, and its visualization can be seen in the Appendix .

Table 6.2: Regression Coefficients

Feature	**Linear Regression Coef**	**Ridge Coef**	**Lasso Coef**
global_avg_norm_price	1.9925	1.9923	1.9926
global_std_norm_price	-0.0833	-0.0832	-0.0833
primary_norm_price	-0.6883	-0.6882	-0.6883
primary_sale_year	-0.0050	-0.0050	-0.0050
second_last_norm_price	-0.4109	-0.4108	-0.4109
second_last_sale_year	-0.0203	-0.0203	-0.0203
third_last_norm_price	-0.5678	-0.5677	-0.5678
third_last_sale_year	-0.0232	-0.0232	-0.0232
fourth_last_norm_price	-0.1734	-0.1733	-0.17334
fourth_last_sale_year	0.0203	0.0203	0.0203
last_sale_year	0.0308	0.0308	0.0308

Table 6.3: Regression Metrics and Parameters

Metrics/Parameters	**Linear Regression**	**Ridge Regression (L2)**	**Lasso Regression (L1)**
alpha	None	0.0514	0.00001
RMSE mean (10-CV)	0.2445	0.2445	0.2445
STD of RMSE	0.0019	0.0018	0.0018
R^2 score	68.7132 $\times 10^{-2}$	68.7133 $\times 10^{-2}$	68.7132 $\times 10^{-2}$

Upon examining Table 6.2, it can be observed that the majority of the coefficients are strikingly similar, barring the coefficients of global_avg_norm_price. However, it is worth noting that the coefficients for all other features are exactly the same between Linear Regression and Lasso Regression. Similarly, Table 6.3 highlights a comparable trend where the R^2 scores of Linear Regression and Lasso Regression are identical, while the Ridge Regression produced slightly superior results. These results can be explained simply by the fact that the optimal alpha value of Lasso Regression is extremely close to zero, which makes it significantly similar to the original simple Linear Regression.

6.3 Multi-layer Perceptron

In the MLP (Multilayer Perceptron) section, the training process of the model is initiated by defining its architecture. This includes specifying the number of input and output neurons, as well as determining the structure of the hidden layers. Additionally, choices regarding activation functions, batch sizes, optimizers, and the number of epochs are made to configure the model.

Once the model is defined, the training process begins. During training, the model iteratively updates its weights and biases using backpropagation and gradient descent. The aim is to minimize the loss function and improve the model's performance.

To evaluate the performance of the model, the validation loss is monitored during training. This loss represents the discrepancy between the predicted values and the actual values in the validation dataset. A plot of the validation loss against the epochs is created, allowing for a visual assessment of the model's learning progress. The minimum validation loss is identified as the point at which the model achieves its best performance on the validation dataset.

In addition to the validation loss, the R^2 score is calculated as a measure of the model's predictive capability.

After extensive experimentation and tuning, we have identified the optimal parameters for the MLP (Multi-Layer Perceptron) model. These parameters were selected based on their impact on model performance (validation loss and R^2 score) and their ability to yield the best results for our specific task.

During the experimentation phase of our MLP (Multi-Layer Perceptron) model, we focused on exploring the impact of the number of hidden layers and the number of neurons in the hidden layer first. As mentioned in Chapter 3.6, it is commonly recommended to use a single hidden layer for regression tasks.

By varying the number of hidden layers and the number of neurons within those layers, we examined the model's performance in terms of the validation loss and the R^2 score. Figure 6.2 illustrates the relationship between the number of hidden layers and the validation loss. We observed that the minimum validation loss was achieved when there was only one hidden layer with 400 neurons. This finding suggests that a single hidden layer with 400 neurons is sufficient for capturing the underlying patterns in the data and minimizing the prediction errors. Similarly, Figure 6.3 displays that the R^2 score reached its maximum value when using a single hidden layer with 400 neurons.

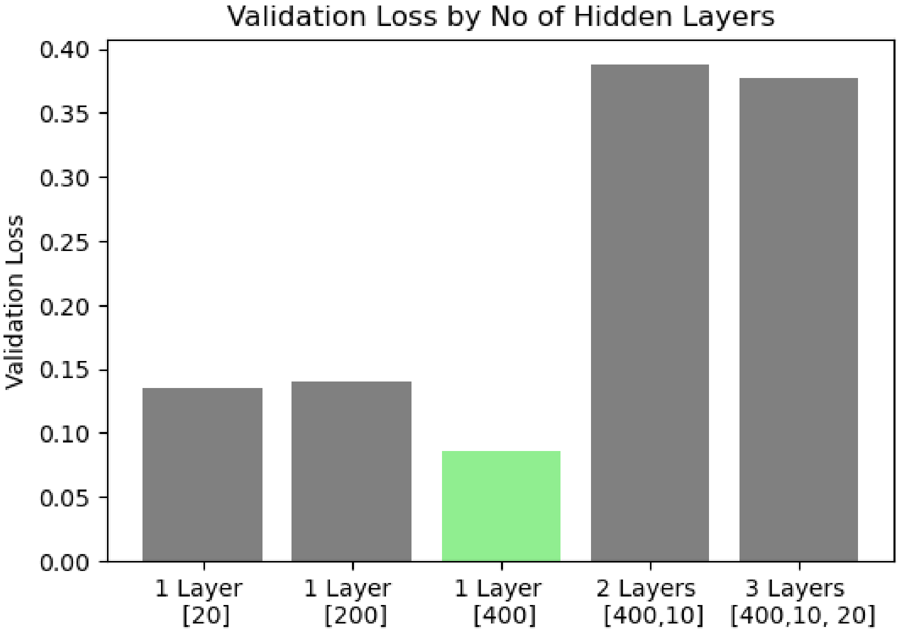


Figure 6.2: Validation Loss by the Number of Hidden Layers

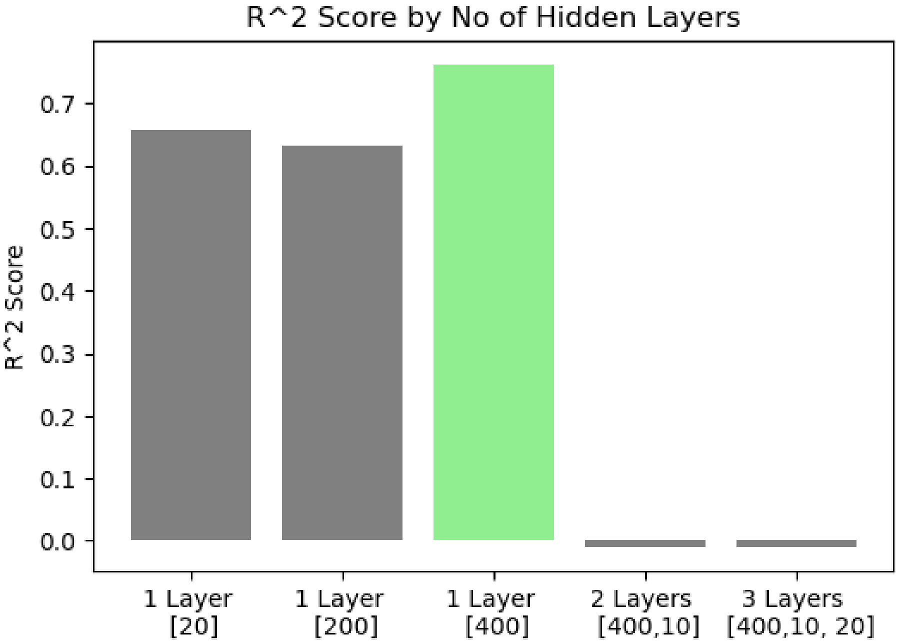


Figure 6.3: R^2 Score by the Number of Hidden Layers

In the subsequent phase of our experimentation, we focused on determining the optimal number of epochs for training the MLP model. We explored a range of epochs, spanning from 100 to 1600, and evaluated the model's performance in terms of the validation loss and the R^2 score.

Figure 6.4 provides insights into the relationship between the number of epochs and the validation loss. It is evident that as the number of epochs increases, the validation loss gradually decreases. However, we observed that after reaching a minimum, the validation loss stabilizes around 1000 epochs.

Figure 6.5 depicts the R^2 score corresponding to different numbers of epochs. We noted that the R2 score steadily increases as the number of epochs increases, reaching its maximum value of 0.79 at 1000 epochs.

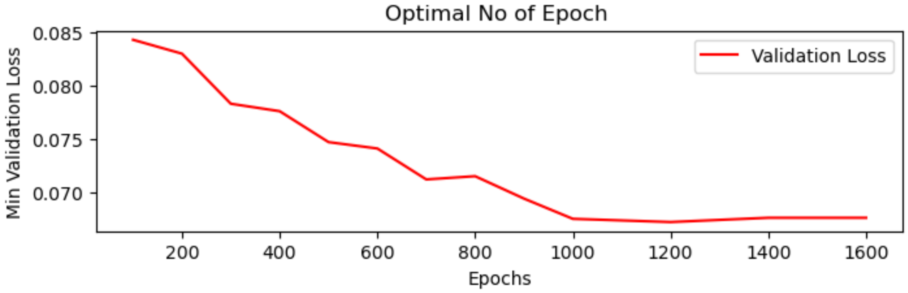


Figure 6.4: Optimal Number of Epochs

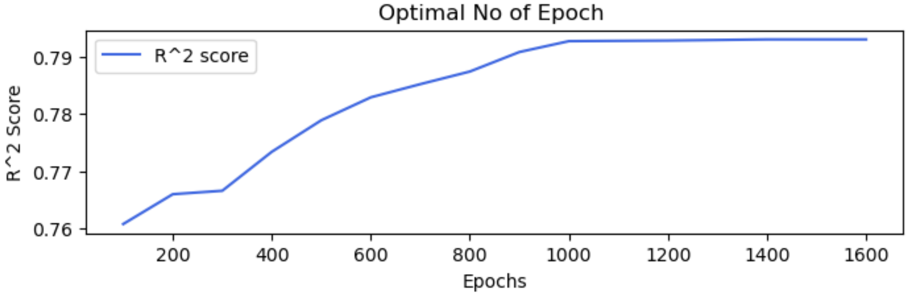


Figure 6.5: Optimal Number of Epochs

In the subsequent stage of our experimentation, we focused on exploring different activation functions for the hidden layer (Layer 1) and the output layer (Layer 2) of the MLP model. Specifically, we investigated the effects of using ReLu, Sigmoid, and Tanh activation functions.

To evaluate the performance of various activation function combinations, we generated heatmaps depicting the validation loss and R^2 score for different combinations of Layer 1 and Layer 2 activation functions. These heatmaps, as shown in Figures 6.6 and 6.7, provide valuable insights into the impact of activation functions on the model's performance. Based on the observed results, it was found that the combination of Tanh activation for the hidden layer (Layer 1) and Sigmoid activation for the output layer (Layer 2) yielded the most favorable outcomes. This combination achieved a validation loss of 0.072 and an R^2 score of 0.79, indicating better performance compared to other activation function combinations.

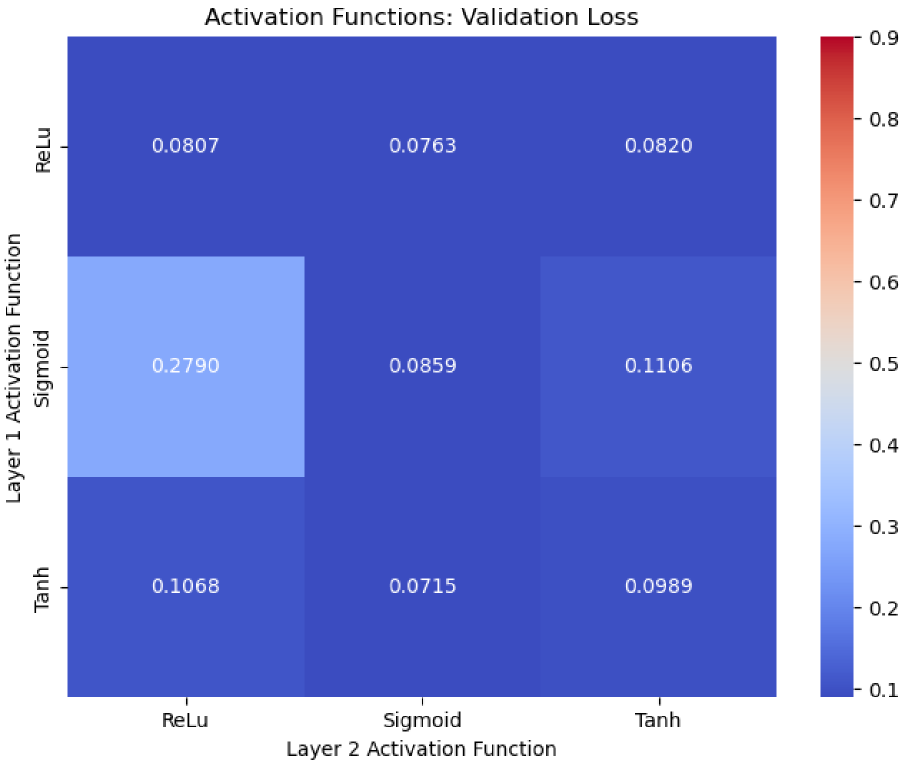


Figure 6.6: Activation Functions Heatmap: Validation Loss

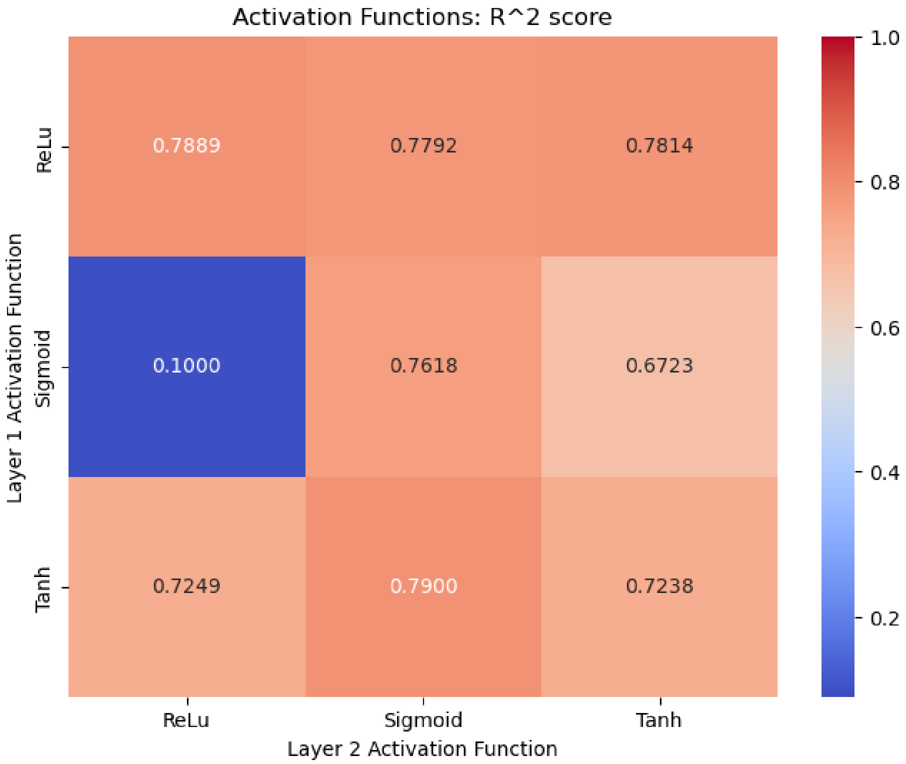


Figure 6.7: Activation Functions Heatmap: R^2 score

Finally, after conducting experiments with different optimization algorithms including Adadrad, SGD, and Adam, we carefully evaluated their performance in minimizing the cost function. Our analysis, as depicted in Figures 6.8, clearly indicates that the Adam optimizer outperformed the other algorithms, exhibiting both the minimal validation loss and the maximum R^2 score. This compelling evidence led us to select the Adam optimizer as the optimal choice for our model in achieving superior performance in the task of minimizing the cost function.

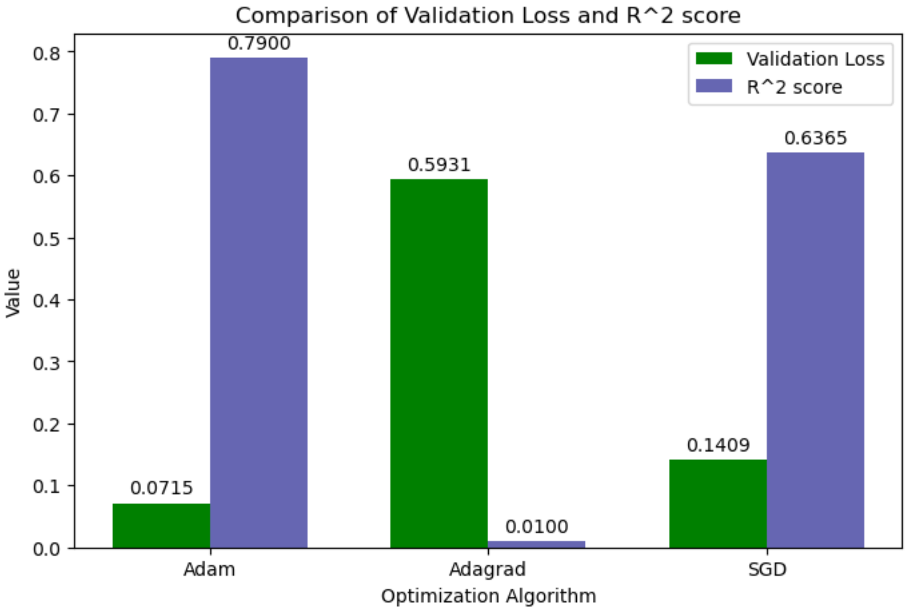


Figure 6.8: Validation Loss and R^2 score

Chapter 7

Conclusion

7.1 Statistical Analysis

The descriptive analysis of the NFT dataset revealed key characteristics of the market. The dataset covered a span of four years and included over 4 million unique NFTs and 6 million trades. The transaction frequency indicated that many NFTs were traded only once, and the Games, Collectible, and Art categories dominated the market in terms of transaction volume. The price distribution exhibited wide variation, with a skewed distribution and notable differences between mean and median prices. Each category displayed its own price distribution pattern, emphasizing the heterogeneity within the NFT market.

The examination of the Gini Index and Lorenz Curve shed light on the inequality of NFT price distributions. The Lorenz Curve indicated a highly unequal distribution, with a small percentage of traders holding a disproportionate share of wealth. The Gini Index further confirmed the high level of price inequality, with Art and Games categories showing the greatest inequality. In contrast, the Utility and Metaverse categories displayed a more equal distribution of prices. These findings provided insights into the concentration of wealth within the NFT market.

The time-series analysis revealed a steady increase in the average price and volume of NFT transactions since 2017, with a peak in early 2021. However, the market experienced periods of fluctuation, indicating volatility. The trends varied across categories, with different categories exhibiting peaks of popularity at different times. Moving average techniques highlighted the overall trends, with Art showing an increasing trend, Collectible demonstrating a decreasing trend, and Other, Games, and Utility categories displaying price fluctuations.

The analysis of correlations between NFT categories identified interesting relationships. Collectibles and Art showed the highest positive correlation in terms of price changes, indicating similarities in their movements. Other, Art, and Utility categories also exhibited moderate positive correlations. In terms of transaction volume, Collectibles, Games, and Other categories showed relatively high correlations with each other, while Art and Metaverse categories had lower correlations. The Utility category displayed low correlations with all other categories. These findings provided insights into the interconnectedness and patterns within the NFT market. The results of the correlation analysis could potentially be insightful for investors that try to recognize patterns and invest in the categories that are potentially correlated. For example, when Metaverse's popularity on the market increases, that could give a hint to investors that they could also start investing in Games based on their correlation.

7.2 Hierarchy of Clusters

The clustering analysis aimed to organize the NFTs into distinct clusters and explore the most prominent ones using the K-means clustering technique. Five main features were selected for the analysis, including local_avg_price, global_avg_price, counts_norm, local_std, and global_std, which provided insights into the price and transaction characteristics of the NFTs. The initial clustering process involved identifying the optimal number of clusters through the "Elbow Method" among NFTs that were sold more than once and resulted in the largest cluster constituting over 76% of the NFTs sold twice. Further analysis revealed that the clusters followed triangular shapes, indicating a linear relationship between the features.

Based on the initial clustering results, the hierarchy of clusters was refined by excluding the larger clusters representing NFTs sold only once or twice. The remaining NFTs were subjected to hierarchical clustering, resulting in four subclusters. The hierarchy of clusters consisted of nine classes, organized at different levels based on feature extraction peculiarities and K-means hierarchical clustering. Each class exhibited distinct characteristics in terms of average price, standard deviation, and count.

Intriguingly, the hierarchical clustering analysis of the NFT dataset yielded surprising insights that shed light on the nature of the market. As the K-means algorithm organized NFTs into distinct clusters based on their sales history and price dynamics, several clusters exhibited unexpected characteristics and trends. Notably, certain clusters displayed unique price behavior and transaction patterns, challenging conventional assumptions about NFT categorization. These surprising findings, uncovered during the clustering process, have the potential to redefine our understanding of the NFT market and its underlying dynamics.

7.3 Price Prediction

In the feature selection and data preprocessing stage, 11 significant features were selected for predicting the last sale price of NFTs. The dataset was preprocessed by grouping NFTs and normalizing their prices. Linear regression, Lasso, and Ridge regularization techniques were applied, and coefficients of the features were calculated. The results showed a similarity between linear regression and Lasso regression coefficients, while Ridge regression performed slightly better.

The Multi-Layer Perceptron (MLP) model was trained by defining its architecture and configuring hidden layers, activation functions, batch sizes, optimizers, and epochs. The model underwent experimentation to determine the optimal number of hidden layers and neurons, as well as the number of epochs. The combination of Tanh activation for the hidden layer and Sigmoid activation for the output layer yielded the best results. The Adam optimizer outperformed other optimization algorithms in minimizing the cost function.

7.4 Contributions

The present study makes significant contributions to the understanding of non-fungible tokens (NFTs) through various analyses and modeling techniques. Firstly, we provide a comprehensive analysis of the NFT market, examining transaction frequency, category preferences, and price distributions. This analysis sheds light on the heterogeneous nature of the NFT ecosystem, offering valuable insights into market dynamics and category-specific trends. We also investigate the inequality of NFT price distributions using the Gini Index and Lorenz Curve revealing significant wealth disparities within the market, highlighting the concentration of prices in certain categories. These insights have implications for understanding market dynamics and informing investment decisions.

Secondly, we investigate the hierarchical clustering of NFTs, organizing them into distinct clusters based on their sales history and price dynamics. By applying K-means clustering and analyzing the resulting clusters, we uncover meaningful patterns and relationships within the NFT market. This hierarchical clustering approach allows for a more granular understanding of the market, enabling stakeholders to identify significant trends and make informed decisions based on the characteristics of each cluster.

Additionally, we develop predictive models for NFT price estimation, incorporating feature selection and utilizing linear regression, regularization techniques (Lasso and Ridge), and a Multi-Layer Perceptron (MLP) model. These models provide valuable tools for price prediction and market analysis, helping investors and collectors navigate the NFT market with greater confidence.

7.5 Limitations and Future work

The research process encountered several noteworthy challenges that should be acknowledged. Firstly, the limitations of the dataset became apparent as it relied on open-source data and accessible APIs. Consequently, the dataset may not have captured a comprehensive representation of all NFT sales, resulting in potential missing data. Furthermore, the dataset is limited in its temporal scope, extending only until April 2021, which restricts the analysis of more recent developments in the NFT market. Another significant challenge arose from the distribution of sales across NFTs, with the majority being sold only once. This limited the availability of historical data required for training the predictive models.

As a result, the dataset's capacity to capture robust patterns and trends may have been compromised due to the scarcity of repeated sales and insufficient historical information.

Moreover, due to the incomplete market picture, capturing and incorporating price fluctuations into the machine-learning models proved to be a challenge. The absence of a comprehensive understanding of market dynamics and the lack of real-time data hindered the ability to construct accurate predictive models that could effectively account for price volatility.

Due to the limitations of data availability and the focus on analyzing existing data, this research utilized datasets collected by a group of scientists using specific APIs. While the available data provided valuable insights into the NFT market, it was restricted to a specific temporal scope and might not have encompassed more recent developments in the rapidly evolving NFT landscape. Consequently, the analysis primarily concentrated on the stable and curated datasets at hand, making it challenging to incorporate later data that could potentially impact the broader understanding of the NFT market dynamics.

These limitations underscore the need for caution when interpreting the findings and emphasize the importance of further research to address data availability issues, expand the temporal scope, and develop methodologies that can account for the unique characteristics of the NFT market.

In terms of future work, there are several areas that warrant further exploration and investigation. Firstly, expanding the dataset by including more comprehensive and up-to-date data will enhance the accuracy and reliability of the analysis. This can be achieved by accessing additional sources of NFT transaction data or partnering with platforms to obtain a more comprehensive view of the market. Furthermore, exploring advanced machine learning techniques and models can improve the accuracy of price prediction and market analysis. Another avenue for future research is the exploration of the impact of regulatory frameworks on the NFT market. As the market continues to grow and attract increased attention, it is crucial to understand how regulatory policies and interventions may shape its dynamics. Investigating the influence of regulatory factors such as intellectual property rights, taxation, and consumer protection measures can provide insights into the potential risks and opportunities associated with NFT investments and transactions.

Bibliography

[1] R. P. Adams. K-means clustering and related algorithms. COS 324 – Elements of Machine Learning, 2017. Retrieved May 6, 2023, from `https://www.codecademy.com/learn/machine-learning/modules/dspath-clustering/cheatsheet`.

[2] R. P. Adams. Hierarchical clustering. Unpublished lecture, COS 324 - Elements of Machine Learning, Princeton University, Princeton, NJ, USA, Nov. 2019.

[3] C. Aggarwal. *Neural Networks and Deep Learning: A Textbook.* Springer, Cham, Switzerland, 2018.

[4] M. Ali and S. Bagui. Introduction to nfts: The future of digital collectibles. *International Journal of Advanced Computer Science and Applications*, 12(10), 2021.

[5] P. Baheti. Activation functions in neural networks, May 2021. Retrieved May 6, 2023, from `https://www.v7labs.com/blog/neural-networks-activation-functions`.

[6] J. C. Bezdek and R. J. Hathaway. Numerical taxonomy with the use of cluster analysis. *Journal of the International Association for Mathematical Geology*, 3(2):169–186, 1971.

[7] P. Bholowalia and A. Kumar. Ebk-means: A clustering technique based on elbow method and k-means in wsn. In *2014 Fifth International Conference on Computational Intelligence, Communication Systems and Networks*, 2014.

[8] R. G. Brown. Smoothing, forecasting and prediction of discrete time series. *Annals of Mathematical Statistics*, 21(4):505–513, 1950.

[9] S. R. Chakravarty. *Inequality, polarization and poverty: Advances in distributional analysis.* Springer Science & Business Media, 2013.

[10] Codecademy. Clustering: K-means [cheatsheet]. Retrieved May 6, 2023, from `https://www.codecademy.com/learn/machine-learning/modules/dspath-clustering/cheatsheet`.

[11] W. Cong and Z. He. Blockchain disruption and smart contracts. *The Review of Financial Studies*, 32(5):1754–1797, May 2019.

[12] Wikipedia contributors. Gini coefficient. `https://en.wikipedia.org/wiki/Gini_coefficient`. Accessed: May 6, 2023.

[13] F. Costa, W. La Cava, and A. Tagarelli. Show me your nft and i tell you how it will perform: Multimodal representation learning for nft selling price prediction. *Knowledge-Based Systems*, 232:106985, 2022.

[14] M. Ezekiel. *Methods of correlation analysis.* Wiley, 1930.

[15] F. A. Farris. The gini index and measures of inequality. *The American Mathematical Monthly*, 117(10):851–864, 2010.

[16] C. F. Gauss. *Theoria Motus Corporum Coelestium.* 1809.

[17] C. Gini. *Variabilità e mutabilità.* Tipografia di Paolo Cuppini, 1912.

[18] A. W. Gumelar and T. H. Pudjiantoro. Nft coin price prediction (non-fungible token) using k-nearest neighbors method. *IOP Conference Series: Materials Science and Engineering*, 1201(1):012014, 2023.

[19] S. Hansun. A new approach of moving average method in time series analysis. *Journal of Information Processing Systems*, 8(2):261–268, Jun. 2012.

[20] J. Harper. Jack dorsey's first ever tweet sells for $2.9m. *BBC News*, March 2021.

[21] J. A. Hartigan and M. A. Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108, 1979.

[22] E. Howcroft. Canadian crypto investor snags virtual real estate plot for record us$2.4 million. *National Post*, Nov. 2021.

[23] S. T. Howell, M. Niessner, and D. Yermack. Initial coin offerings: Financing growth with cryptocurrency token sales. *The Review of Financial Studies*, 33(9):3925–3974, Sep. 2020.

[24] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning: with Applications in R.* Springer, New York, 2013.

[25] Roman Kräussl and Antonio Tugnetti. Non-fungible tokens (nfts): A review of pricing determinants, applications and opportunities. *Journal of Economic Behavior Organization*, 197:27–49, 2022.

[26] J. Legendre and C. A. L. Bassi. *Nouvelles méthodes pour la détermination des orbites des comètes.* 1805.

[27] R. G. Lorenz. Methods of measuring the concentration of wealth. *Publications of the American Statistical Association*, 9(70):209–219, 1905.

[28] M. Nadini, L. Alessandretti, F. Di Giacinto, M. Martino, G. Luca, and A. Baronchelli. Mapping the nft revolution: Market trends, trade networks, and visual features. *IEEE Access*, 9:89322–89333, 2021.

[29] Matthieu Nadini. Non fungible tokens dataset. [Online]. Available: `https://osf.io/wsnzr/?view_only=319a53cf1bf542bbbe538aba37916537`, 2020.

[30] F. Nielsen. Hierarchical clustering. In D. E. Holmes, editor, *Introduction to HPC with MPI for Data Science*, pages 195–211. Springer International Publishing, 2016.

[31] M. Nielsen. *Neural Networks and Deep Learning.* Determination Press, 2015.

[32] K. Pearson. Mathematical contributions to the theory of evolution. iii. regression, heredity and panmixia. *Philosophical Transactions of the Royal Society of London*, 187:253–318, 1896.

[33] S. Reyburn. Jpg file sells for $69 million, as 'nft mania' gathers pace. *The New York Times*, Mar. 2021.

[34] F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408, 1958.

[35] M. Rosenfeld. Overview of colored coins, 2012.

[36] Q. Wang, R. Li, Q. Wang, and S. Chen. Non-fungible token (nft): Overview, evaluation, opportunities and challenges. *Journal of Ambient Intelligence and Humanized Computing*, 12(9):8951–8966, 2021.

[37] Z. J. Zhang. Cryptopricing: Where comes the value for cryptocurrencies and nfts? *Int. J. Res. Marketing*, 40(1):22–29, Mar. 2023.

[38] Z. Zheng, S. Xie, H.-N. Dai, X. Chen, and H. Wang. Blockchain challenges and opportunities: a survey. *Int. J. Web Grid Serv.*, 14(4):352–375, 2018.