# Análise de Características Estruturais de Tokens não Fungíveis no Ethereum

**Samuel de Oliveira Ribeiro**[1], **Dayan Ramos Gomes**[1]
**Emanuel Coutinho**[2], **Glauber Dias Gonçalves**[1]

[1]Universidade Federal do Piauí (UFPI) - CSHNB

[2]Universidade Federal do Ceará (UFC) - Campus Quixadá

{samueloliveira0014, dayanramos5537}@gmail.com

emanuel.coutinho@ufc.br, ggoncalves@ufpi.edu.br

***Resumo.*** *Token não fungível ou NFT é um objeto digital insubstituível por qualquer outro objeto, seja do mesmo tipo ou valor, com atributos que provam a sua propriedade a uma pessoa ou organização via redes blockchain. A indústria de artes e mídias digitais vem adotando gradativamente NFTs devido a sua segurança para definir autoria, transferência, royalties desses tokens, entre outros recursos que podem ser programados em contratos inteligentes. Como NFT é uma tecnologia nova com popularidade em ascendência, há oportunidades para desenvolvimento de ferramentas que auxiliem os usuários no consumo desse tipo de token. Neste trabalho, realizamos uma análise e caracterização de coleções de NFTs baseada em dados extraídos do OpenSea, que é a maior plataforma de comercialização de NFTs na atualidade. Utilizamos uma abordagem de classificação não supervisionada para conhecer propriedades estruturais dessas coleções. Isso nos permitiu definir quatro classes de coleções de NFTs que podem ser facilmente compreendidas por usuários para facilitar o comércio e a valoração de seus tokens.*

***Abstract.*** *Non-fungible token or NFT is a digital object that cannot be replaced by any other object, whether of the same type or value, with features that prove its ownership to a person or organization via blockchain networks. The arts and digital media industry have gradually adopted NFTs due to their security for defining authorship, transfer, and royalties of these tokens, among others, that can be programmed in smart contracts. As NFT is a new technology with increasing popularity, there are opportunities for the development of tools that assist users in consuming this type of object. In this work, we conducted an analysis and characterization of NFT collections based on data extracted from OpenSea, which is the largest NFT trading platform currently. We used an unsupervised classification approach to learn about the structural properties of these collections. It allowed us to define four classes of NFT collections that can be easily understood by users to facilitate the trade and valuation of their tokens.*

## 1. Introdução

Token não fungível ou NFT, do inglês *Non-fungible Token*, é um objeto digital registrado em plataformas blockchain, tipicamente, associado a conteúdo de texto ou imagem, que

o confere características únicas e o torna também um objeto colecionável. Adicionalmente, NFT permite definição de um autor (criador), transferências de propriedade entre usuários, *royalties* para o criador, dentre outros recursos do ambiente distribuído de blockchains. Devido a essas características, NFTs vem sendo adotados por artistas para criação e distribuição de conteúdo digital, visando a proteção do direito autoral e do ganho com royalties na revenda de itens.

A plataforma blockchain Ethereum é a pioneira na emissão de NFTs e que concentra a vasta maioria de coleções desses *tokens* atualmente. Ela oferece padrões de programação, i.e., contratos inteligentes, específicos para esse tipo de *token*. A maioria desses padrões estabelecem um único contrato para uma coleção de NFTs. Por sua vez, cada NFT da coleção é vinculado de forma imutável a atributos que o tornam único. Esses atributos incluem em especial um link para um arquivo de mídia temático (e.g., jogos, artes, vídeos ou redes sociais, identificadores do autor e proprietário, valores de venda e *royalties* para o autor. Toda transferência de propriedade, i.e., venda, de NFT é automaticamente registrada como uma transação na blockchain. Propriedades de segurança das blockchains garantem descentralização, transparência e imutabilidade na transação, corroborando para a credibilidade no comércio de NFTs [Revoredo 2021].

Nesse contexto, NFTs abrem um novo caminho para utilização e veiculação de objetos digitais, i.e., *tokens*, sob a Internet, onde o aspecto mais relevante é a autoria ou propriedade do *token*. Por exemplo, em 11 de março de 2021, o artista Beeple realizou a venda de sua obra de arte digital em formato de NFT na blockchain Ethereum pelo valor de US$ 69 milhões [Christie's 2021]. Em 22 de março de 2021, o fundador do Twitter Jack Dorsey vendeu o NFT do seu famoso primeiro tweet pelo valor de US$ 2,9 milhões [Okonkwo 2021]. Essas obras podem ser acessadas gratuitamente na Internet e facilmente replicadas. Contudo, quanto mais popular e copiado na Internet é o *token*, mais benefícios ele pode trazer ao seu proprietário, que possui direitos exclusivos sobre a sua comercialização e imagem.

Uma plataforma de comercialização ou mercado de NFTs é um ambiente virtual onde são oferecidos vários serviços e facilidades para divulgação e vendas desses *tokens*. *OpenSea* é uma das principais plataformas de comercialização de NFTs, e vem experimentando um crescimento acelerado desde o seu lançamento em 2018[1]. A plataforma já negociou milhões de dólares em NFTs, e a procura por esses *tokens* cresce diariamente. Para os usuários interessados neles, é importante entender a classificação dos NFTs nas plataformas antes de realizar uma compra. A classificação pode incluir informações sobre a autenticidade do NFT, a sua raridade, tipo de mídia (e.g., arte digital, memes, músicas, fotografias), entre outros aspectos relevantes.

Contudo, grandes plataformas como *OpenSea* não divulgam publicamente critérios e métodos adotados para classificação de NFT[2]. A falta de transparência pode comprometer a confiança dos autores na plataforma, os levando à especula sobre privilégios ou preferências a autores ou temas determinados. Logo, é importante a adoção de métodos alternativos de classificação, que priorizem a transparência e a padronização das informações apresentadas aos usuários.

[1]https://opensea.io

[2]Até a escrita desse artigo, o *OpenSea* reporta 4 categorias: arte digital, itens de jogos, colecionáveis e memes em seu website, sem informar como NFTs são enquadrados nelas.

A maioria das propostas da literatura foca na classificação de usuários com base em características extraídas de suas transações na blockchain ([Norvill et al. 2017, Wang et al. 2020, Valadares et al. 2021, Aspembitova et al. 2021, Wu et al. 2021]). Outras linhas de trabalhos focam em classificações de tráfego malicioso na rede ([Xu et al. 2020, Rebello et al. 2020, Hu et al. 2021]). Por sua vez, análises de características de *tokens* e transações no Ethereum são demonstrados com propósito o de uma análise de desempenho da plataforma em [Oliveira et al. 2021, Singh and Hafid 2019]. Com o foco especificamente em NFTs, [Casale-Brunet et al. 2021] realizaram análises de características dos usuários de NFTs utilizando modelos de grafos a partir das transações de compra e venda entre os usuários do comércio de NFTs. Contudo, nenhum desses trabalhos foca em métodos para identificar características estruturais relevantes de coleções de NFTs que permitam uma classificação dessas de forma transparente para os usuários desse ecossistema.

Este artigo tem o objetivo de preencher essa lacuna da literatura com duas contribuições principais:

- Identificação de características estruturais, i.e., atributos imutáveis codificados no contrato, assim como estatísticas de comercialização das coleções de NFTs importantes para o entendimento do mercado e a valoração desses *tokens*.
- Proposta de uma metodologia alternativa para classificar coleções de NFT baseado nas características acima referidas.

Nesse sentido, realizamos a análise e caracterização de coleções de NFTs na plataforma *OpenSea*, baseada em 5 atributos relacionados à estrutura e às transações dos *tokens* na própria plataforma. Em nosso estudo foram coletados dados de 50.000[3], mais as top 100 mais populares, coleções de NFTs na plataforma *OpenSea*, realizada seleção das características mais relevantes e identificadas classes de coleções de NFTs a partir dessas características. Acreditamos que a metodologia de seleção de características e classificação proposta neste trabalho possibilitam um melhor entendimento sobre coleções de NFTs e os seus fins comerciais.

As próximas seções deste artigo possuem a seguinte organização. Na Seção 2, discutimos os trabalhos relacionados. A metodologia de coleta de dados e o seu tratamento são descritos na Seção 3. A Seção 4 apresenta com mais detalhes os resultados obtidos, mostrando o método utilizado para o agrupamento dos dados e a caracterização dos dados. Finalmente, a Seção 5 expõe as considerações finais.

## 2. Trabalhos Relacionados

A maioria dos trabalhos relacionados foca em classificações de usuários da plataforma Ethereum baseado em características extraídas de suas transações na blockchain. Em [Norvill et al. 2017], os autores focaram na identificação da finalidade de uma conta de usuário a partir de características obtidas no contrato inteligente dessa mesma conta. Já em [Wang et al. 2020] é feita a extração de características das contas a partir da associação com contas públicas já rotuladas na API Etherscan. [Valadares et al. 2021] propuseram técnicas para classificação de usuários como comuns ou profissionais com base em suas transações realizadas na plataforma Ethereum extraídas a partir da API Etherscan.

[3]A plataforma não especifica qual o critério de resposta da API, logo não sabemos se as 50.000 coleções foram as mais antigas, recentes ou com algum critério de aleatoriedade entre todas.

Em [Aspembitova et al. 2021] é feita uma análise de comportamento dos usuários no mercado de criptoativos, em particular, a estrutura comportamental dos usuários que realizam transações de compra e venda de criptoativos. Foi utilizada uma combinação de algoritmos de clusterização e classificação para identificar padrões comportamentais dos usuários. Os autores de [Wu et al. 2021] propuseram um modelo de categorização de usuários do Ethereum, por meio de algoritmos de agrupamento. Este trabalho também fará uso de extração de características das transações, contudo com foco em classificar coleções de NFTs. Estes trabalhos utilizam algoritmos de aprendizado não supervisionado para agrupar as contas dos usuários de acordo com suas características, segundo o objetivo de cada conta.

Outro grupo de trabalhos trata de modelos de classificação no contexto de segurança do Ethereum. Em [Xu et al. 2020], foi desenvolvido um modelo de classificação com *random forest* para diferenciar o tráfego malicioso em um tipo de "ataque de eclipse", este ataque permite que um ator mal-intencionado possa interferir com os nós de uma rede. Como o nome sugere, o ataque visa ofuscar a visão de um participante da rede P2P, para provocar interrupções gerais ou com o objetivo de preparação para ataques mais sofisticados. Os autores de [Rebello et al. 2020] aplicaram modelos de aprendizado de máquina supervisionados para identificar lavagem de dinheiro na rede Bitcoin. Foi desenvolvida uma rede neural LSTM por [Hu et al. 2021] para identificar tipos de contrato inteligente mais comuns no Ethereum, incluindo a classe de contratos suspeitos relacionados a contravenções como comércio e jogos ilegais. Em nosso trabalho aplicamos técnicas de clusterização de dados para agrupar coleções de NFTs baseadas em suas características.

Em [Oliveira et al. 2021] foi desenvolvido métodos de previsão de falha no processamento de contratos por mineradores do Ethereum utilizando classificadores com aprendizado supervisionado. Por sua vez, [Singh and Hafid 2019] desenvolveram modelos de ML para prever o tempo de confirmação de uma transação, explorando o impacto de classes de dados desbalanceadas no treinamento e teste dos modelos selecionados. Essas pesquisas possuem um propósito de análise de desempenho da plataforma, já o nosso projeto busca propor uma metodologia alternativa para classificar coleções de NFT baseada nas suas características.

Em [Casale-Brunet et al. 2021] foi feita uma análise de comportamento dos usuários que comercializavam as top oito coleções de NFTs mais populares da plataforma *OpenSea*. Os autores coletaram dados de transações ERC-721 e criaram uma rede que mostrava as conexões entre os *tokens*. Eles então aplicaram várias medidas de centralidade e detecção de comunidades para identificar os *tokens* mais importantes e as comunidades mais distintas dentro da rede. Os dados utilizados foram coletados a partir da Blockchain Ethereum, utilizando uma API pública. Consistiram em cerca de 2 milhões de transações relacionadas a 8 coleções de NFTs. Nosso propósito é diferente, focamos em características estruturais das coleções dos NFTs. Neste trabalho proposto é feito o processamento de cerca de 50.000 coleções de NFTs, e como metodologia de análise foi utilizada uma abordagem de clusterização, baseado em características como total de vendas, total de itens na coleção e a quantidade de atributos em cada coleção.

## 3. Dados e Metodologia

Nesta seção, apresentamos nossos conjuntos de dados e metodologia de processamento adotada. Primeiro, introduzimos o processo de coleta de dados e o volume do conjunto de dados obtido, a seguir descrevemos as etapas de processamento desses dados. Finalmente, explicamos como realizamos a seleção de características estruturais de coleções de NFTs.

### 3.1. Coleta dos dados

Utilizamos a API *OpenSea*[4] para coletar informações de coleções de NFTs da plataforma blockchain Ethereum. *OpenSea* é uma das maiores e mais populares plataformas de comercialização de NFTs no momento da escrita deste trabalho. Essa plataforma oferece uma API REST para obter dados das NFTs nela comercializadas. A seguir, descrevemos o processo de coleta com essa API.

Desenvolvemos um programa na linguagem Python 3 para realizar requisições HTTP a *endpoints* gratuitos da API REST *OpenSea*[5]. A API *OpenSea* oferece uma ampla variedade de *endpoints*, cada um dos quais fornece um conjunto específico de dados sobre os itens da plataforma, como preços, histórico de transações, informações de propriedade e outros. Por exemplo, para coletar informações sobre uma determinada NFT, é possível usar o *endpoint /asset/*{*contract-address*}*/*{*token-id*}. Este *endpoint* retornará uma grande quantidade de informações relevantes sobre o item, incluindo o preço atual, a histórico de transações e o proprietário atual.

**Tabela 1. *Endpoints* da *OpenSea* utilizados para coletar os dados.**

| Código | *Endpoint* | Parâmetros |
|---|---|---|
| endpoint-1 | /api/v1/collections?offset=offset&limit=limit | offset, limit |
| endpoint-2 | /api/v1/collection/collection-name | collection-name |
| endpoint-3 | /api/v1/collection/collection-name/stats | collection-name |

A Tabela 1 lista os três *endpoints* utilizados para coletar os dados desse trabalho. O *endpoint*-1 é responsável por fornecer uma lista de todas as coleções suportadas e examinadas pelo *OpenSea*. O *endpoint*-2 é utilizado para recuperar informações mais detalhadas sobre uma coleção individual, incluindo lista de editores, informações sobre conta de pagamentos, informações descritivas da coleção e atributos desta coleção e muitos outros dados. O *endpoint*-3 foi utilizado para buscar estatísticas de cada coleção específica, incluindo dados de precificação da coleção em tempo real.

Dessa forma nosso programa foi configurado para realizar requisições a cada 1 segundo no *endpoint*-1 e 0.5 segundos nos *endpoint*-2 e *endpoint*-3. Estes tempos foram definidos com base em testes de estresse realizados na API, utilizamos então os valores de tempo mínimos para que não houvesse bloqueios de requisições. Ao final da coleta foram totalizadas 166 requisições no *endpoint*-1 e 50.000 requisições para cada um dos demais *endpoints*. Para cada requisição feita foi salvo um arquivo JSON contendo a resposta da requisição.

[4]https://opensea.io/

[5]Tentamos o contato com o *OpenSea* para obter a chave que dá direito a explorar outros *endpoints* privados da plataforma mas até o momento não fomos atendidos.

Tendo em vista que o *endpoint*-1 possui uma limitação nos seus parâmetros, ficamos restritos a baixar dados de 50.000 (cinquenta mil) coleções distintas. Após a realização das requisições do *endpoint*-1, as requisições do *endpoint*-2 e *endpoint*-3 foram realizadas individualmente de forma automatizada para cada coleção. Adicionalmente, coletamos dados das 100 coleções de NFTs mais populares de acordo o ranqueamento mantido pela plataforma *OpenSea*[6]. Ao fim desse processo conseguimos dados de um total de 49.325 coleções distintas da plataforma. Esses dados foram salvos no formato JSON estruturados com as informações requeridas para o processamento a seguir.

### 3.2. Processamento dos dados

O objetivo desta etapa é processar os dados das coleções de NFT coletados no formato JSON, original da API *OpenSea*. A Tabela 2 sumariza os dados que foram filtrados e organizados após essa etapa de processamento inicial. Observa-se que para cada coleção obtivemos um total de 21 conjuntos de dados descritos em cada linha da Tabela 2.

**Tabela 2. Conjunto de características coletadas para cada coleção de NFTs.**

| Característica | Descrição |
|---|---|
| *slug* | Identificação única e legível por URL de um item ou coleção na plataforma *OpenSea*. |
| *total-volume* | Total de volume de vendas em ETH na plataforma *OpenSea*. |
| *total-sales* | Total de vendas na plataforma *OpenSea* em ETH. |
| *total-supply* | Representa o número total de NFTs que foram criados para uma coleção. |
| *count* | Representa o número atual de NFTs disponíveis para venda. |
| *num-owners* | Número de proprietários de um item na plataforma *OpenSea*. |
| *average-price* | Preço médio de venda dos itens na plataforma *OpenSea*. |
| *num-reports* | Número de relatórios de abuso associados a um item na plataforma *OpenSea*. |
| *market-cap* | Capitalização de mercado de uma determinada coleção na plataforma *OpenSea*. |
| *qtd-traits* | Número de características ou atributos associados a um item ou coleção em particular na plataforma *OpenSea*. |
| *is-nsfw* | Indicação se o conteúdo de um item ou coleção é considerado "não seguro para o trabalho". |
| *is-rarity-enabled* | Indicação se o sistema de raridade está ativado para uma coleção específica na plataforma *OpenSea*. |

O próximo passo do processamento de dados foi realizar uma análise estatística dos valores das características para todas as coleções de forma a selecionar aqueles com maior teor de informações para o estudo. Assim, removemos características que não apresentavam variações entre seus valores para a maioria das coleções de NFTs. Logo, características como o *'is-rarity-enabled'* e *'is-nsfw'* não foram utilizadas nas análises visto que mais de 99% das coleções apresentavam os valores padrões da plataforma. A Tabela 3 mostra os conjuntos de dados selecionados e algumas informações estatísticas sobre os valores desses dados.

[6]https://opensea.io/rankings?sortBy=total_volume, acessado em 19 de janeiro de 2023.

**Tabela 3. Dados estatísticos das coleções de NFT considerando todas as 49325 coleções coletadas.**

| | mean | std | min | max |
|---|---|---|---|---|
| *total-volume* | 126 | 7707 | 0 | 1105317 |
| *total-sales* | 49 | 1526 | 0 | 208389 |
| *total-supply* | 72 | 1290 | 0 | 113393 |
| *count* | 72 | 1290 | 0 | 113393 |
| *num-owners* | 26 | 548 | 0 | 65608 |
| *average-price* | 0.008 | 0.319 | 0 | 49.282 |
| *num-reports* | 0.001 | 0.061 | 0 | 5 |
| *market-cap* | 63 | 4124 | 0 | 753078 |
| *qtd-traits* | 0.556 | 2.171 | 0 | 217 |

A seguir, realizamos uma análise preliminar dos valores das características pré selecionadas para todas as coleções de NFTs buscando possíveis inconsistências. Observamos que uma parte relevante das coleções tinham indícios de não se tratar de coleções profissionais ou comerciais. Isso pôde ser constatado, em especial, em coleções com volume de transações (característica "*total-volume*") com valor zero ou quantidade de *tokens* (característica "*count*") igual a zero. Dessa forma, optamos por remover todos os conjuntos de coleções suspeitas, ou seja, coleções de testes ou usadas para propósito de pirataria ou contravenções[7], visando garantir que as coleções incluídas no estudo tenham liquidez e sejam relevantes para os investidores e colecionadores. A Tabela 4 mostra informações descritivas do total de 1157 coleções de NFTs[8], resultantes dessa etapa.

**Tabela 4. Dados estatísticos das 1157 coleções de NFT selecionadas para esse estudo após a análise de consistência dos dados.**

| | mean | std | min | max |
|---|---|---|---|---|
| *total-volume* | 5419 | 50097 | 3.7e-14 | 1105317 |
| *total-sales* | 2083 | 9759 | 1 | 208389 |
| *total-supply* | 1922 | 7518 | 1 | 113393 |
| *count* | 1922 | 7518 | 1 | 113393 |
| *num-owners* | 750 | 2742 | 1 | 573110 |
| *average-price* | 0.342 | 2.061 | 0 | 49 |
| *num-reports* | 0.077 | 0.397 | 0 | 5 |
| *market-cap* | 2728 | 26826 | 0 | 753078 |
| *qtd-traits* | 2.467 | 8.009 | 0 | 217 |

### 3.3. Extração e Seleção de Características

Nessa seção, descrevemos a etapa de seleção e extração de características de NFTs após o processamento dos dados coletados. Nesse sentido, calculamos a correlação

[7]Foi realizado uma análise qualitativa de uma amostra desse grupo de coleções considerado "suspeitos". Foi visto que a maioria era utilizado como uma ferramenta para disseminar pirataria de mídias, como a transmissão de jogos de futebol e filmes.

[8]Note que o valor mínimo em *total-volume*, apesar de muito próximo de zero, não foi descartado. As coleções com esses valores baixos usualmente são utilizadas em testes na rede, onde o valor de venda dos *tokens* é pequeno para minimizar os gastos totais.

entre todas as características mostradas na Tabela 4 com o intuito de identificar características similares que poderiam não acrescentar informação para a nossa classificação na seção seguinte. Para quantificar as correlações utilizamos o coeficiente de Pearson [Bussab and Morettin 2017].

Outra técnica que foi utilizada é a aplicação de escala logarítmica nos dados coletados. Isso é útil para reduzir o impacto dos altos contrastes e dispersões causados por valores extremos nas características (i.e., *outliers*). Ao usar uma escala logarítmica, é possível visualizar melhor a distribuição dos dados e fazer comparações mais precisas entre as coleções. A partir desses dados processados já conseguimos analisar a correlação entre as variáveis, definir subconjuntos e avaliar o agrupamento destes dados.

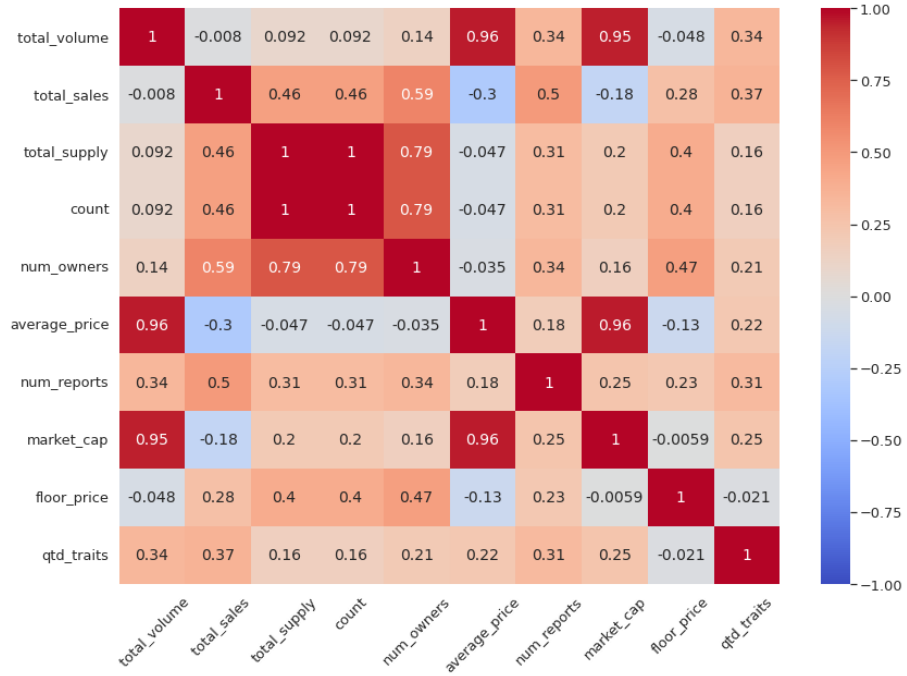


**Figura 1. Correlação de Pearson aplicada aos dados filtrados e em escala logarítmica.**

A Figura 1 apresenta a matriz de correlação de Pearson aplicada às coleções de NFT. A correlação de Pearson mede a relação linear entre duas variáveis e seus valores variam de -1 a 1, onde -1 representa uma correlação negativa perfeita, 0 uma ausência de correlação e 1 uma correlação positiva perfeita. Observando a Figura, podemos verificar que as características estão moderadamente relacionadas entre si. Algumas das correlações mais significativas são:

- *total-volume* e *average-price*: Essas características possuem uma correlação positiva muito forte de 0.956846. Isso significa que, quanto maior o volume de *tokens* negociados, maior tende a ser o preço médio das transações.
- *total-volume* e *market-cap*: Estas duas características possuem uma correlação positiva muito forte de 0.951453. Isso significa que quanto maior for a capitalização de mercado, maior tende a ser o volume de *tokens* negociados.
- *total-supply* e *count*: Estas duas características estão relacionadas a partir de uma forte correlação positiva de 1.0. Isso significa que o número total de *tokens* disponíveis e o número de *tokens* gerados são diretamente proporcionais.

- *total-supply* e *num-owners*: Estas duas características também estão relacionadas a partir de uma correlação positiva de 0.788841. Isso significa que o número total de *tokens* disponíveis tende a influenciar o número de proprietários de *tokens*, ou seja, quanto maior o número de *tokens* disponíveis, maior tende a ser o número de proprietários.

É importante observar que a correlação não implica causalidade, ou seja, não se pode concluir que há uma relação de causa e efeito entre essas variáveis apenas com base em suas correlações, e outros fatores não incluídos neste estudo podem estar influenciando as variáveis de interesse, como por exemplo a popularidade do criador da coleção, a data de lançamento da coleção, a qualidade dos *tokens* e a estratégia de marketing utilizada para a comercialização.

Existem duas abordagens principais de seleção de recursos não supervisionados: filtro e *wrapper* [Aspembitova et al. 2021]. A abordagem de filtro seleciona os recursos mais relevantes com base em critérios específicos, como correlação, variância e entropia. Por outro lado, a abordagem de *wrapper* define primeiro subconjuntos de recursos e, em seguida, avalia-os com base em um algoritmo de clusterização. Nessa seção adotamos ambas as abordagens, começamos calculando a correlação entre as várias características e com base nos resultados da correlação sugerimos os seguintes 3 conjuntos de características para análises específicas:

- Conjunto-1: todas as características.
- Conjunto-2: todas as características, removendo as que apresentaram alta correlação (i.e., coeficiente de $Pearson \geq 0.75$). Neste conjunto foram descartadas as características *'count'*, *'num-owners'*, *'average-price'*, e *'market-cap'*.
- Conjunto-3: Contendo apenas as características não correlacionadas com nenhuma outra característica (i.e., coeficiente., de $Pearson < 0.75$). Neste conjunto restaram apenas as características *'num-reports'*,*'floor-price'* e *'qtd-traits'*.

Os três conjuntos de características identificados serão utilizados para alcançar nossos resultados quanto à identificação de classes de coleções de NFTs na seção a seguir.

## 4. Resultados

Nesta seção, discutimos classes de coleções de NFTs utilizando os dados coletados e processados na seção anterior. Primeiramente, mostramos nossa abordagem para definir a quantidade de classes de NFTs. A seguir, discutiremos as características estruturais para cada uma dessas classes suportado por análises qualitativas e quantitativas.

### 4.1. Classes de NFTs

Nosso primeiro resultado consiste em definir classes para o total de 1157 coleções de NFTs coletados e selecionados de acordo com os passos da seção anterior. Tal classificação é importante pois visa oferecer ferramentas para plataformas de comercialização de NFTs exporem seus itens de forma mais adequada aos usuários compradores e proprietários. Por outro lado, a classificação pode ajudar aos artistas e criadores de coleções de NFTs diferenciarem seus produtos, o que impacta na definição de preços para esses, i.e., valoração dos *tokens*. Essa classificação é desafiante pois necessita ser baseada em critérios claros que convençam atores do ecossistema de NFTs a adotar as classes propostas. Contudo, observamos que a prática usual de classificação é baseada em

critérios definidos unicamente pelas plataformas de comercialização. Conjecturamos que esses critérios são subjetivos e de interesse particular da plataforma, por não haver transparência da metodologia e transparência dos critérios adotados para os artistas, criadores e proprietários de NFTs.

Tendo em vista que os nossos dados não são rotulados, utilizamos o classificação não supervisionada via o algoritmo K-means [Ahmad and Mohammed 2019]. Neste caso, seguimos a recomendação de [Aspembitova et al. 2021] que recomenda esse algoritmo para classificação do comportamento de usuários em plataformas blockchain. Os autores deste trabalho fazem essa recomendação após analisarem outros algoritmos alternativos para classificação como DBScan e Optics. Contudo, antes de utilizar K-means é necessário determinar o número adequado de classes, i.e., *clusters*. Para isso foi utilizado o método do cotovelo (também conhecido como *elbow method*), que funciona traçando um gráfico que mostra a variação da soma dos erros quadráticos em relação ao número de *clusters*. O objetivo é encontrar o ponto no gráfico onde a adição de mais um *cluster* não resulta em uma redução significativa na soma dos erros quadráticos. Esse ponto é chamado de "cotovelo" e representa o número ideal de *clusters* para o conjunto de dados em questão.

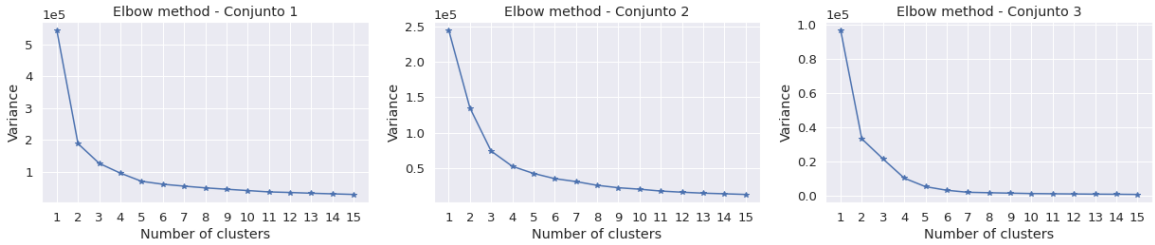


**Figura 2. Método de cotovelo aplicado aos 3 conjuntos de dados.**

A Figura 2 apresenta os resultados da aplicação do método do cotovelo aos três conjuntos de coleções definidos na seção anterior, onde o eixo x representa o número de classes (*"clusters"*) avaliados, ao passo que o eixo y representa a variância da soma dos erros quadráticos. Observando a Figura 2, podemos inferir que o número de *cluster* ideal seria entre 3 e 5 nos 3 conjuntos de dados, visto que esses pontos são onde visivelmente houve uma maior queda na curva. Contudo, essa análise unilateral não é precisa para definir o número k de *clusters*. Além da avaliação visual do método do cotovelo, existem outras métricas que podem ser usadas para avaliar a qualidade do modelo K-means e determinar o número ideal de *clusters*. Algumas dessas métricas que utilizamos nesse trabalho incluem:

- Coeficiente de *Silhouette*: essa métrica mede a similaridade entre cada ponto e os outros pontos dentro do mesmo *cluster*, bem como a dissimilaridade em relação aos pontos dos outros *clusters*. O coeficiente varia de -1 a 1, com valores mais próximos de 1 indicando *clusters* mais bem definidos e valores mais próximos de -1 indicando que os pontos podem ter sido atribuídos a *clusters* errados.
- Índice de Calinski-Harabasz: esse índice mede a relação entre a variação entre os *clusters* e a variação dentro dos *clusters*. Valores mais altos do índice indicam que os *clusters* são mais bem definidos e têm menos sobreposição.
- Índice Davies-Bouldin: esse índice mede a relação entre a dissimilaridade média

de cada *cluster* e a dissimilaridade entre os *clusters*. Valores mais baixos do índice indicam que os *clusters* são mais bem definidos e têm menos sobreposição.

A Figura 3 mostra o resultado das métricas analisadas para cada conjunto de dados. A melhor escolha de K é aquela que maximiza o Coeficiente de *Silhouette* e o Índice de Calinski-Harabasz e minimiza o Índice Davies-Bouldin.

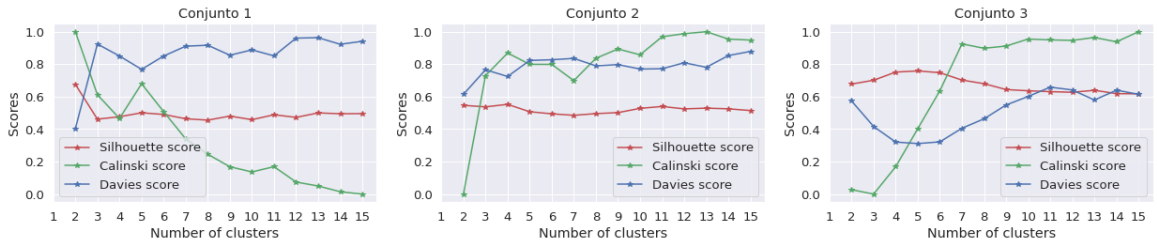


**Figura 3. Resultado das métricas Coeficiente de *Silhouette*, Índice de Calinski-Harabasz e Índice Davies-Bouldin aplicadas aos 3 conjuntos de dados.**

No Conjunto 1, o melhor K é 5, pois apresentou o maior Coeficiente de *Silhouette* e Índice de Calinski-Harabasz e um valor razoável para o Índice Davies-Bouldin. No Conjunto 2, o melhor K é 4, pois apresentou o maior Coeficiente de *Silhouette* e Índice de Calinski-Harabasz até 10 *clusters* e o menor valor para o Índice Davies-Bouldin. Por fim, para o conjunto 3, K=5 é o melhor número de *clusters* para esse conjunto de dados. Isso é indicado pelo alto valor do índice de Calinski-Harabasz e a baixa pontuação do índice Davies-Bouldin nesse valor de K, bem como pelo Coeficiente de *Silhouette* relativamente alto.

Para prosseguir nas análises iremos utilizar os dados do conjunto-2, visto que ele foi o grupo que melhor obteve o valor K observando as métricas utilizadas e é o grupo que possui número intermediário de atributos, tendo descartado apenas 4 atributos com alta correlação.

### 4.2. Características Estruturais das Classes

Nessa seção focamos nas características estruturais para cada uma das classes, i.e., *clusters*, encontradas na seção anterior, tomando como referência o conjunto-2 com 4 *clusters* e 5 características. A Tabela 5 apresenta dados referentes à média aritmética de cada característica dos *clusters* obtidos a partir do agrupamento de coleções de NFTs. A seguir, discutimos as características de cada *cluster* suportada pelos dados quantitativos da Tabela 5:

- *Cluster* 0: Tem um volume de transações relativamente baixo em comparação com os outros *clusters*. No entanto, as vendas totais são bastante significativas

**Tabela 5. Valor médio das características para cada *cluster*.**

| *Cluster* | *total-volume* | *total-sales* | *total-supply* | *num-reports* | *qtd-traits* |
|---|---|---|---|---|---|
| 0 | 13.40 | 246.75 | 1347.60 | 0.00 | 5.04 |
| 1 | 9.26 | 74.42 | 580.70 | 0.00 | 0.00 |
| 2 | 0.00 | 314.37 | 644.59 | 0.00 | 0.70 |
| 3 | 61885.75 | 22064.33 | 13414.64 | 0.88 | 11.32 |

(i.e., média de US$ 246.75). Isso sugere que as transações desse *cluster* envolvem itens de alto valor. O número de relatórios (característica *num-reports*) é zero, o que pode indicar que a comunidade que negocia neste *cluster* é menos ativa em termos de relatar problemas. A quantidade média de características (i.e., característica *qtd-traits*) é intermediária, aproximadamente 5.04, sugerindo que este *cluster* pode ter alguns *traits* distintos, mas não é tão segmentado quanto o *cluster* 3.

- *Cluster* 1: Os membros deste *cluster* possuem uma quantidade total de *tokens* (característica *total-supply*) relativamente baixa de NFTs (média de 580.70 *tokens*), vendas e possuem 0 *traits*. Isso pode indicar que esses NFTs não são tão distintos ou interessantes para os compradores em comparação com os outros *clusters*, e também que a comunidade que negocia neste *cluster* é relativamente pequena e pode não ter uma demanda significativa por NFTs.
- *Cluster* 2: Este *cluster* tem o menor volume total entre os *clusters*, mas possui um número relativamente alto de vendas totais (i.e, 314.37) em comparação com outros *clusters*, o que indica que, apesar do baixo volume, os dados agrupados nesse *cluster* têm um alto valor de transações. A quantidade média de características é baixa (i.e, 0.70), sugerindo que este *cluster* pode não ter características únicas.
- *Cluster* 3: Este *cluster* tem o maior volume total, bem como o maior número de vendas e proprietários em comparação com os outros *clusters*. A quantidade média de *traits* é a mais alta (média de 11.32), sugerindo que este *cluster* possui características mais distintas e únicas em comparação com os outros. Além disso, este *cluster* tem um número significativo de relatórios (média de 0.88), indicando uma forte atividade em torno dele.

No geral, as características médias de cada *cluster* podem ajudar a entender melhor a natureza dos NFTs em cada grupo e informar decisões de investimento ou comercialização. Por exemplo, um investidor pode optar por focar em NFTs do *cluster* 2, que possuem um valor total de vendas relativamente alto, mas baixo volume total, enquanto pode decidir evitar NFTs do *cluster* 3 devido ao número médio de relatórios negativos que eles receberam.

Por sua vez, a nomenclatura dos *clusters* pode ser definida com base em suas características qualitativas, i.e., um nome que reflita o tipo de coleções de NFTs que eles contêm. Algumas sugestões de nomes para cada *cluster*, com base nas observações feitas anteriormente, podem ser:

- *Cluster* 0: “Coleções Diversificadas” ou “Misto”, pois possuem uma média moderada em todas as características.
- *Cluster* 1: “Coleções de Baixo Valor” ou “Iniciantes”, pois possuem valores baixos em todas as características.
- *Cluster* 2: “Coleções de Alta Venda” ou “Vendáveis”, pois possuem uma média alta em vendas totais.
- *Cluster* 3: “Coleções de Alto Valor” ou *“Premium”*, pois possuem médias altas em todas as características, especialmente na característica *num-reports*.

Esses nomes destacam as características distintas de cada *cluster* e também sugerem o tipo de coleção de NFTs que pode ser encontrada em cada um deles. De forma comparativa, a plataforma *OpenSea* possui classes próprias para as suas coleções de NFTs, divididas por categorias temáticas e objetivas, como:

- Arte digital: Obras de arte digital únicas, incluindo pinturas, desenhos, animações e outros tipos de mídia.
- Itens de jogos: Itens virtuais em jogos de computador ou *mobile*, como espadas, armaduras, *skins*, personagens e outros tipos de objetos.
- Colecionáveis: coleções de NFTs que representam objetos físicos ou digitais raros e colecionáveis, como cartões de beisebol, figurinhas, moedas e outros itens.
- Memes: coleções de NFTs que representam memes populares da internet, como o "Nyan Cat" e "Bad Luck Brian", que foram transformados em arte digital colecionável.

Essas são apenas algumas das categorias de coleções de NFTs disponíveis na plataforma *OpenSea*. Assim como nos *clusters* definidos neste estudo, cada categoria da *OpenSea* tem sua própria comunidade de compradores e vendedores, com seus próprios valores e tendências de mercado. Contudo, essa plataforma não é transparente para seus usuários a metodologia utilizada para incluir coleções nas categorias existentes, visto que ao criar coleções utilizando a plataforma, não existe uma opção para que o usuário categorize sua criação. E também existem diversas coleções que não possuem uma categoria definida e não são encontradas pelo filtro de categoria padrão já disponibilizado na *OpenSea*.

## 5. Considerações Finais

NFTs são ativos digitais únicos que são criados e registrados em uma blockchain, e que estão em bastante ascendência no mercado global. Neste trabalho, propomos uma metodologia para identificar características estruturais relevantes de coleções de NFTs, e a partir dessas características utilizamos uma abordagem para classificar essas coleções. Para realizar esse trabalho, foram coletados dados de 50.000 coleções de NFTs, mais as top 100 coleções de NFTs mais populares, na plataforma *OpenSea*. A partir desses dados foram realizados uma série de procedimentos para remoção de coleções inconsistentes, para na sequência identificarmos características estruturais relevantes para a classificação dessas coleções. Acreditamos que a metodologia de seleção de características e agrupamentos proposta neste trabalho possibilitam um melhor entendimento sobre coleções de NFTs e os seus fins de comércio.

Para trabalhos futuros, planejamos o desenvolvimento e construção de um classificador automático supervisionado a partir dos *clusters* encontrados e continuar a coleta de dados utilizando APIs de terceiros como o *OpenSea*. Adicionalmente, planejamos construir uma infraestrutura de nó na rede e API própria para coletar vários dados da Blockchain Ethereum e outras, não apenas se referindo a NFTs mas também outros tipos de contratos e *tokens*.

## Referências

Ahmad, T. and Mohammed, H. (2019). K-means clustering algorithm: Applications in data science and bioinformatics. *Big Data Analytics and Computational Intelligence*, 1(1):1–13.

Aspembitova, A. T., Feng, L., and Chew, L. Y. (2021). Behavioral structure of users in cryptocurrency market. *PLOS ONE*, 16(1):1–19.

Bussab, W. d. O. and Morettin, P. A. (2017). *Estatística Básica*. Saraiva.

Casale-Brunet, S., Ribeca, P., Doyle, P., and Mattavelli, M. (2021). Networks of ethereum non-fungible tokens: A graph-based analysis of the erc-721 ecosystem. https://arxiv.org/abs/2110.12545. arXiv.

Christie's (2021). Beeple (b. 1981). https://onlineonly.christies.com/s/first-open-beeple/beeple-b-1981-1/112924.

Hu, T., Liu, X., Chen, T., Zhang, X., Huang, X., Niu, W., Lu, J., Zhou, K., and Liu, Y. (2021). Transaction-based classification and detection approach for ethereum smart contract. *Information Processing Management*, 58(2):102462.

Norvill, R., Fiz Pontiveros, B. B., State, R., Awan, I., and Cullen, A. (2017). Automated labeling of unknown contracts in ethereum. In *2017 26th International Conference on Computer Communication and Networks (ICCCN)*, pages 1–6.

Okonkwo, I. E. (2021). Nft, copyright; and intellectual property commercialisation. *SSRN*. https://ssrn.com/abstract=3856154.

Oliveira, V. C., Almeida Valadares, J., A. Sousa, J. E., Borges Vieira, A., Bernardino, H. S., Moraes Villela, S., and Dias Goncalves, G. (2021). Analyzing transaction confirmation in ethereum using machine learning techniques. *SIGMETRICS Perform. Eval. Rev.*, 48(4):12–15.

Rebello, G., Hu, Y., Thilakarathna, K., Batista, G., Seneviratne, A., and Duarte, O. C. (2020). Melhorando a acurácia da detecção de lavagem de dinheiro na rede bitcoin. In *Anais do XXXVIII Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*, pages 728–741, Porto Alegre, RS, Brasil. SBC.

Revoredo, T. (2021). Nfts e sua sofisticação nos blockchains. https://exame.com/blog/tatiana-revoredo/nfts-e-sua-sofisticacao-nos-blockchains/.

Singh, H. J. and Hafid, A. S. (2019). Transaction confirmation time prediction in ethereum blockchain using machine learning. https://arxiv.org/abs/1911.11592. arXiv.

Valadares, J., Oliveira, V., Sousa, J., Bernardino, H., Villela, S., Vieira, A., and Gonçalves, G. (2021). Identificação de perfis de comportamento de usuários no ethereum utilizando técnicas de aprendizado de máquina. In *Anais do IV Workshop em Blockchain: Teoria, Tecnologias e Aplicações*, pages 60–73, Porto Alegre, RS, Brasil. SBC.

Wang, M., Ichijo, H., and Xiao, B. (2020). Cryptocurrency address clustering and labeling. https://arxiv.org/abs/2003.13399. arXiv.

Wu, S. X., Wu, Z., Chen, S., Li, G., and Zhang, S. (2021). Community detection in blockchain social networks. *J. Commun. Inf. Networks*, 6:59–71.

Xu, G., Guo, B., Su, C., Zheng, X., Liang, K., Wong, D. S., and Wang, H. (2020). Am i eclipsed? a smart detector of eclipse attacks for ethereum. *Computers Security*, 88:101604.