



**Universidade Federal de Uberlândia
Faculdade de Matemática**

Bacharelado em Estatística

**Métodos Multivariados aplicados na seleção
de fundos imobiliários**

Bruno Picioli Toselli

**Uberlândia-MG
2023**

Bruno Picioli Toselli

**Métodos Multivariados aplicados na seleção
de fundos imobiliários**

Trabalho de conclusão de curso apresentado à Co-
ordenação do Curso de Bacharelado em Estatística
como requisito parcial para obtenção do grau de
Bacharel em Estatística.

Orientador: Leandro Alves Pereira

Uberlândia-MG

2023



**Universidade Federal de Uberlândia
Faculdade de Matemática**

Coordenação do Curso de Bacharelado em Estatística

A banca examinadora, conforme abaixo assinado, certifica a adequação deste trabalho de conclusão de curso para obtenção do grau de Bacharel em Estatística.

Uberlândia, _____ de _____ de 20_____

BANCA EXAMINADORA

Leandro Alves Pereira

Lúcio Borges de Araújo

Marcelo Tavares

**Uberlândia-MG
2023**

AGRADECIMENTOS

Primeiramente agradeço a Deus, por dar força, saúde, conhecimentos e por sempre está comigo em realizar meus objetivos.

Agradeço a minha mãe Silvia Picioli e ao meu pai Fernando Toselli, que sempre me apoiaram as minhas decisões e me ofereceram todo o suporte para chegar até aqui. São duas pessoas importantes na minha vida, pois me ensinaram a ser batalhador e nunca desistir do próprio sonho.

Um agradecimento especial para minha namorada Flávia, obrigado por tudo, você me mostrou o que realmente é ser uma pessoa determinada e guerreira para trilhar o próprio caminho. Tenho muita admiração por você. Agradeço também aos meus amigos que fiz na faculdade, sem eles a graduação não seria fácil.

Agradeço ao meu orientador, Prof. Leandro Alves Pereira, por toda a compreensão, confiança e conselhos. Pela administração das aulas que você ministrou, você foi um exemplo notável e um professor excepcional.

Agradeço a todos os professores que lecionaram disciplinas para mim, com certeza sem vocês eu não chegaria até aqui.

E a Universidade Federal de Uberlândia, por todo acolhimento e ensinamentos passados para minha vida inteira.

RESUMO

Investir em bens de valor para obter retornos futuros é uma prática comum na sociedade brasileira, incluindo imóveis, poupanças e previdências privadas. De acordo com a Associação Brasileira das Entidades dos Mercados Financeiro e de Capitais, os brasileiros estão cada vez mais interessados no mercado financeiro de renda variável. O mercado financeiro é composto por transações de compra e venda de ativos, como fundos imobiliários, realizadas por instituições como a Bolsa de Valores. Este estudo tem como objetivo melhorar o entendimento das características dos fundos imobiliários e auxiliar investidores na seleção desses ativos. Para isso, foi utilizada a técnica de análise de componentes principais para reduzir a dimensionalidade dos dados e o método de agrupamentos K-means para identificar grupos de fundos semelhantes. Os resultados obtidos mostraram que os fundos imobiliários podem ser agrupados em sete clusters com perfis diferentes. Esses agrupamentos podem ser utilizados por investidores para selecionar os fundos mais adequados à sua estratégia de investimento.

Palavras-chave: Investimentos, Mercado Financeiro, Renda Variável, Estatística Multivariada.

ABSTRACT

The purpose of this study is to improve the understanding of the characteristics of real estate funds and assist investors in the selection of these assets. Investing in value assets for future returns is a common practice in Brazilian society, including real estate, savings, and private pensions. According to the Brazilian Association of Financial and Capital Market Entities, Brazilians are increasingly interested in the stock market. The financial market is composed of buy and sell transactions of assets such as real estate funds, performed by institutions such as the Stock Exchange. This study used the technique of principal component analysis to reduce the dimensionality of the data and the K-means clustering method to identify groups of similar funds. The results showed that real estate funds can be grouped into seven clusters with different profiles. These groups can be used by investors to select the most appropriate funds for their investment strategy.

Keywords: Investment, Stock Market, Multivariate Statistics..

SUMÁRIO

Lista de Figuras	I
Lista de Tabelas	II
1 Introdução	1
2 Metodologia	3
2.1 Base de Dados	3
2.2 Índices Financeiros	3
2.3 Análise de Componentes Principais (PCA)	6
2.4 Análise de agrupamentos	7
2.4.1 Medidas de Dissimilaridade	7
2.4.2 Agrupamentos Não Hierárquicos e Método K-Means	8
2.4.3 Critérios para seleção do número de clusters	9
3 Resultados	10
3.1 Análise descritiva dos Fundos Imobiliários	10
3.2 Análise das Componentes Principais	11
3.3 Algoritmo K-means	13
4 Conclusões	17
Referências Bibliográficas	18
Apêndice A Apêndice	19

LISTA DE FIGURAS

3.1	Gráfico de barras da relação dos fundos imobiliários por setor	10
3.2	Gráfico do cotovelo para seleção de clusters.	14

LISTA DE TABELAS

3.1	Tabela com estatísticas descritivas índices financeiros	11
3.2	Tabela dos autovalores com percentuais da variação para cada componente . . .	11
3.3	Tabela dos coeficientes para cada componentes e variáveis	12
3.4	Tabela Soma de Quadrados entre Cluster e R-quadrado de 4 a 10 grupos	15
3.5	Tabela de Frequência de fundos imobiliários em cada cluster	15
3.6	Tabela dos clusters com baixa frequência de fundos	16
3.7	Tabela dos grupos com alta frequência de fundos	16
A.1	Tabela de fundo imobiliários coletados	19

1. INTRODUÇÃO

O hábito de investir em algum bem de valor para obter retornos futuros está enraizado na sociedade brasileira, antes da década de 60 os brasileiros investiam principalmente em imóveis físicos ou até mesmo em poupanças, já que na época a situação econômica somada a legislações limitava o desenvolvimento de um mercado financeiro e de capitais ativo [6].

Por outro lado, atualmente houve um aumento significativo de investidores dentro do mercado financeiro e de capitais. Segundo a ANBIMA (Associação Brasileira das entidades dos Mercados financeiro e de Capitais), a sociedade abdicou-se de investimentos tradicionais e passou a usufruir mais de produtos financeiros, cerca de cinquenta e três por cento. Tendo em vista, os brasileiros deixaram a forma tradicional de investimento, migrando para o mercado financeiro e de capitais [7].

O mercado financeiro e de capitais é uma organização onde ocorre compra e venda de bens de valores, tal como ações, commodities, moedas nacionais, fundos de investimentos e imobiliários e outros bens e títulos. O objetivo deste ambiente de negócios é conectar quem necessita de recursos com quem quer investir, sejam pessoas físicas ou jurídicas [3].

No Brasil, o mercado financeiro é controlado por autoridades monetárias, com apoio de instituições governamentais para garantir o cumprimento da política monetária nacional. Essas entidades constituem o sistema financeiro, de forma a organizar e garantir a segurança das transações entre os investidores e captadores [3].

Ainda mais, as transações entre os investidores do mercado financeiro são promovidas por instituições intermediárias, as principais são : Banco Central, caixas econômicas, bancos privados e bolsa de valores [3].

A bolsa de valores é uma empresa, que atualmente no Brasil é chamada de B3.SA, cuja função é reunir investidores através de um sistema de negociações de valores que seja confiável e seguro, permitindo o contato entre os emissores de títulos e bens (empresas e fundos) e uma ampla gama de investidores. [1]

Além de promover o contato entre as empresas e investidores, a B3 é responsável por aspectos essenciais para o correto funcionamento do mercado de capitais, tal como garantir a segurança das negociações, obrigar empresas e os emissores de títulos atuarem de forma clara e correta e facilitar o sistema de transação dos ativos [1].

Além disso, os ativos negociados na bolsa de valores ocorrem por meio de dois sistemas de negociações, mercado primário e secundário. O mercado primário é a relação direta do emissor com o investidor por meio da oferta pública inicial (IPO). Já o mercado secundário realiza-se a

compra e venda dos ativos entre investidores. Dentre esses ativos estão principalmente, ETFs, ações de empresas e fundos imobiliários [11].

Os fundos imobiliários foram introduzidos no Brasil em 1993 pela Lei 8668/93 e regulamentado pela CVM (Comissão de Valores Mobiliários). Apesar dos mais de vinte anos da introdução dos fundos no mercado financeiro brasileiro e ter passado diversas leis de regulamentação, os fundos imobiliários deram-se visibilidade para investidores por volta do ano 2000, por conta da insenção de imposto de renda sobre os rendimentos [9].

Os fundos de investimentos imobiliários é uma comunhão de recursos destinados a aplicação em ativos relacionados ao mercado imobiliário. Geridos por profissionais da área imobiliária com intuito de investir em empreendimentos imobiliários, a fim de ganho de capital e rentabilidade. Ainda mais, cabe ao administrador constituir um fundo e realizar o processo de captação dos recursos através da venda de cotas.

Além disso, os recursos captado pelas cotas poderão ser utilizadas para aquisição dos mais variados setores de imóveis tal como shoppings, hospitais, lajes corporativas, galpões logísticos e títulos de valores mobiliários. Assim, os rendimentos obtidos pela aquisição desses bens são distribuídos periodicamente aos cotistas (investidores) [10].

O investidor ao adquirir uma cota do ativo, torna-se sócio do fundo, tendo participação nos lucros e sendo representado proporcionalmente à sua participação calculado através dos recursos investidos. Caso o investidor decida vender suas cotas, somente poderá recorrer através do mercado secundário da bolsa de valores.

Inúmeros parâmetros são consideráveis no momento de selecionar um fundo imobiliário para investir, assim originando um bom retorno com um risco amenizado. Dentre tantos métodos de seleção de ativos, é natural que o investidor tenha dúvidas no processo de compra desses fundos listados na bolsa de valores.

No mercado financeiro, a seleção errônea de ativos pode ocasionar prejuízos consideráveis no bolso do investidor. Segundo Baroni e Bastos (2018), fundos imobiliários estão relacionados com grandes metrópoles, portanto aumenta o desafio de analisar tais fundos pois as cidades evoluem com o passar do tempo. Assim sendo, o investidor deve priorizar de critérios técnicos e descritivos na seleção de um fundo para amortização de riscos e ganho de rentabilidade.[9]

Em função do métodos de escolhas dos fundos imobiliários muitas informações surgem através das mídias digitais, porém pelo excesso acabam gerando dúvidas no processo de obtenção. Dada a importância do tema, serão utilizados dados históricos e importantes índices de fundos imobiliários retirados do portal Economática para melhor tomada de decisão.

O objetivo desse estudo é apresentar um método de seleção e análise de fundos imobiliários, utilizando técnicas de estatística multivariada, como a análise de componentes principais e o método de agrupamentos, para compreender, agrupar e identificar os grupos de fundos mais semelhantes aos objetivos de cada investidor e, assim, auxiliar na tomada de decisão de seleção de ativos.

2. METODOLOGIA

2.1 BASE DE DADOS

A obtenção da base de dados para fundos imobiliários foi um processo longo e trabalhoso neste trabalho. A princípio, o conjunto de dados seria tratado e calculado através de uma biblioteca em Python chamada Yfinance, onde capta e armazena informações do site Yahoo Finance [5] dentro do pacote, entretanto houve perda de informações na hora de captar os dados. Portanto, descartou-se a possibilidade de obter o conjunto dos índices por esse meio.

Por fim, a base de dados consolidada foi construída a partir do site Economática, uma fonte confiável de dados do mercado financeiro. Com o e-mail do aluno da universidade e acesso à internet da instituição, é possível obter uma variedade de índices financeiros através da plataforma de dados Economática[2].

No portal Economática, foi coletada a série histórica de cada índice financeiro desde o lançamento de fundos imobiliários (IPO) até Dezembro de 2022. Os dados foram armazenados em arquivos CSV para serem tratados e manipulados para obtenção do conjunto de dados final. Utilizou-se o software R [4] para realizar as manipulações e tratamentos dos dados. A base de dados consiste em médias históricas de 8 índices financeiros para cada investimento.

Tratamento e manipulação de dados é uma etapa crucial na análise de dados. É necessário limpar e preparar os dados antes de começar a analisá-los. Com o uso de 8 variáveis, é importante verificar se há valores ausentes ou duplicados e tratá-los. Além disso, é importante verificar a consistência dos dados, como por exemplo, se há valores fora do intervalo esperado. Em seguida, é necessário transformar os dados para que possam ser analisados. Isso pode incluir a padronização dos dados, a criação de novas variáveis e a agregação de dados. Finalmente, é importante verificar a qualidade dos dados antes de iniciar a análise, para garantir que os resultados sejam precisos e confiáveis. O tratamento e manipulação dos dados gerou uma base com 107 fundos imobiliários apresentados na seção de apêndices. De acordo com a base construída foi necessário retirar fundos com valores nulos e na análise foi preciso padronizar as variáveis para manter as escalas no mesmo nível.

2.2 ÍNDICES FINANCEIROS

Indicadores financeiros são medidas quantitativas que apresentam a situação financeira de um fundo de investimentos imobiliário com base nos resultados do período. Esses índices

ajudam como suporte para a tomada de decisão do investidor.

Neste trabalho, foram selecionados 8 indicadores frequentemente utilizados em [9] e [10] para análise e serão descritos a seguir.

Média Histórica do Preço: O preço refere-se as cotas que são negociadas na bolsa, são uma das sinalizações para compra de um fundo, porém analisá-los unicamente não é uma decisão recomendável [9]. A média histórica do preço é diária e referente a data inicial do fundo até Dezembro de 2022. A seguir, é representado a fórmula de cálculo em 2.1.

$$\bar{P} = \frac{\sum_{i=1}^n P_i}{n} \quad (2.1)$$

- \bar{P} : Média Histórica dos preços diários dos fundos imobiliários;
- P_i : Preço diários dos fundos;
- n : Referente a quantidade de dias desde o lançamento do investimento.

Média Histórica do Patrimônio Líquido : Este índice apresenta o que os investidores têm de fato, pois ele é a somatória de tudo que o fundo possui, composto pelo capital subscrito e lucros acumulados, deduzidos dos custos para emissões das cotas e eventuais prejuízos passados. Com isso, esse índice demonstra que quanto maior o valor, mais consistência no mercado [9]. A média histórica do patrimônio é referente aos demonstrativos financeiros trimestrais desde a data de lançamento até Dezembro de 2022. A equação 2.2 representa a fórmula de cálculo.

$$\overline{PL} = \frac{\sum_{i=1}^n PL_i}{n} \quad (2.2)$$

- \overline{PL} : Média Histórica do Patrimônio Líquido trimestral dos fundos imobiliários;
- PL_i : Patrimônio Líquido dos fundos em um determinado trimestre;
- n : Referente a quantidade de trimestres desde o lançamento do investimento.

Média Histórica valor patrimonial por cota (VPC) : Esse indicador financeiro é o valor obtido mediante a divisão do valor do patrimônio líquido pela quantidade de cotas no mercado, representado na fórmula 2.3. O VPC da cota demonstra o quanto do patrimônio do fundo cada cota representa [10]. A média histórica do VPC é diária e referente a data inicial do fundo até Dezembro de 2022. A seguir, é representada a fórmula de cálculo em 2.4.

$$VPC = \frac{PL}{\text{Quantidade de cotas}} \quad (2.3)$$

$$\overline{VPC} = \frac{\sum_{i=1}^n VPC_i}{n} \quad (2.4)$$

- \overline{VPC} : Média Histórica do VPC diários dos fundos imobiliários;
- VPC_i : VPC diários dos fundos;

- n : Referente a quantidade de dias desde o lançamento do investimento.

Média Histórica do preço pelo valor patrimonial por cota (P/VPC) : O P/VPC é um indicador que demonstra o preço de uma ação dividido pelo valor patrimonial por cota correspondente a ela, sendo esse o indicador que diz o quanto os investidores estão dispostos a pagar pelo patrimônio líquido da empresa. Esse indicador é importante, pois um valor acima de 1 representa que o fundo está sendo negociado mais caro na bolsa[10]. A média histórica do preço sobre VPC é diária e referente a data inicial do fundo até Dezembro de 2022. A equação 2.5 representa a fórmula de cálculo.

$$\overline{PVPC} = \frac{\sum_{i=1}^n PVPC_i}{n} \quad (2.5)$$

- \overline{PVPC} : Média Histórica do preço por VPC diários dos fundos imobiliários;
- $PVPC_i$: Preço sobre VPC diários dos fundos;
- n : Referente a quantidade de dias desde o lançamento do investimento.

Média Histórica do Volume : O volume é liquidez dos fundos e indica o processo de compra e venda de um ativo. Ou seja, é a capacidade do investidor de vender a ação no momento. Assim, um valor alto refere-se que o ativo é muito negociado no mercado [10]. Esse índice é referente a média mensal desde a data inicial do fundo até o mês de Dezembro/2022. A equação 2.6 representa a fórmula do indicador.

$$\overline{V} = \frac{\sum_{i=1}^n V_i}{n} \quad (2.6)$$

- \overline{V} : Média Histórica do volume mensal dos fundos imobiliários;
- V_i : Volume diários dos fundos;
- n : Referente a quantidade de dias desde o lançamento do investimento.

Média Histórica dos dividendos : Esse índice apresenta a renda distribuída pelos fundos imobiliário [9]. Para este trabalho foi feita a média dos dividendos mensais desde o IPO . O período é referente desde o IPO até Dezembro/2022. A equação 2.7 representado a forma de cálculo.

$$\overline{D} = \frac{\sum_{i=1}^n D_i}{n} \quad (2.7)$$

- \overline{D} : Média dos dividendos mensais dos fundos;
- D_i : Dividendos mensais dos fundos;
- n : número de meses desde o IPO.

Média Histórica do Retorno : Esta variável é calculada a partir dos preços diários de cada fundo imobiliário e dividido pelo período analisado mais os dividendos, indicando um percentual de rentabilidade em relação ao dia anterior. É calculado subtraindo o preço do dia atual com o preço do dia anterior e dividido pelo preço do dia anterior [2]. O período referente ao indicador é referente ao retorno médio mensal desde o IPO até Dezembro/2022. A equação 2.8 representa a forma de cálculo do índice.

$$R = \frac{P_i - P_{i-1} + D}{P_{i-1}} \quad (2.8)$$

- R : Retorno do fundo;
- P_{i-1} : Preço do dia anterior do fundo;
- P_i : Preço do dia atual do fundo.

Média Histórica do Risco (Volatilidade) : O risco do ativo é calculado através do desvio padrão dos retornos diários de cada fundo e multiplicado pela raiz quadrada dos períodos por ano [2]. Esse indicador serve para o investidor analisar o quanto o ativo oscila em determinado período. O indicador é referente ao risco médio mensal desde a data de lançamento do fundo até Dezembro de 2022. A equação 2.9 representa a fórmula de cálculo.

$$\overline{Risco} = \sigma_{R_i} * \sqrt{n} \quad (2.9)$$

- \overline{Risco} : Risco médio mensal do ativo;
- R_i : Retorno mensal do fundo;
- n : Quantidade de dias desde o lançamento do fundo.

2.3 ANÁLISE DE COMPONENTES PRINCIPAIS (PCA)

A metodologia apresentada é baseada no livro da Sueli Mingoti [12]. Introduzida por Karl Pearson em 1901, a técnica de Análise de Componentes Principais (PCA) tem como principal objetivo de explicar a estrutura de variância e covariância de um vetor aleatório $X' = (X_1, X_2, \dots, X_p)$ onde p são variáveis aleatórias.

O vetor aleatório é construído através das combinações lineares das variáveis originais. Essas combinações lineares são chamadas de componentes principais e são não correlacionadas entre si. Por exemplo, dado um conjunto de p variáveis (X_1, X_2, \dots, X_p) contidas no vetor aleatório $X' = (X_1, X_2, \dots, X_p)$. O primeiro componente principal é dado como $Y_1 = b_{11}X_1 + b_{12}X_2 + \dots + b_{p1}X_p$. Desse modo, se temos p variáveis originais é possível obter p componentes principais.

A técnica PCA geralmente é utilizada para a redução de dimensionalidade das variáveis originais e interpretada a partir das combinações lineares construídas. Desta forma, a informação contida nas p -variáveis originais é substituída pela informação contida em k ($k < p$) componentes não correlacionadas.

O processo de obtenção das componentes principais envolve a decomposição da matriz de covariância ou matriz de correlação caso as variáveis originais tenham diferentes escalas, assim apresentando a necessidade de padronização.

O processo de obtenção das componentes principais via matriz de correlação é estimada pela matriz de covariância amostral. Os autovalores da matriz amostral são as variâncias amostrais dos componentes principais. Os autovetores são os coeficientes associados ao autovalores que serão inseridos nas combinações lineares das componentes principais. Por exemplo, sejam os autovalores $\hat{\lambda}_1, \hat{\lambda}_2, \dots, \hat{\lambda}_p$ da matriz amostral e sejam $\hat{e}_1, \hat{e}_2, \dots, \hat{e}_p$ os autovetores. Então, a n -ésima componente principal é definida por :

$$\hat{Y}_n = \hat{e}'_p Z \quad (2.10)$$

onde Z são as variáveis padronizadas, Z_1, \dots, Z_p e são expressas por $Z_n = \frac{X_n - \mu_n}{\sqrt{\sigma_{nn}}}$, com $n = 1, \dots, p$.

O grau de influência que cada variável Z_n tem sobre a componente Y_n é determinada pela correlação, definida por :

$$r_{Y_n, Z_n} = r_{Y_n, X_n} = e_{ni} \sqrt{\hat{\lambda}_n} \quad (2.11)$$

. A contribuição do n -ésimo componente na explicação total pode ser expressa por :

$$Contr(Y_n) = \frac{\lambda_n}{\sum_{n=1}^p \lambda_n} \quad (2.12)$$

Existem várias maneiras de escolher o número de componentes principais a serem utilizados na análise, os critérios mais comuns são: o critério de Kaiser, o diagrama dos autovalores e o critério da porcentagem da variância acumulada. O critério de Kaiser estabelece que o número de componentes principais deve ser o mesmo número de autovalores que são maiores ou iguais ao valor médio das variâncias das variáveis examinadas. Quando a análise é feita com base na matriz de correlação (variáveis normalizadas), este critério significa excluir componentes com autovalores menores que 1.

Uma vez determinadas as componentes principais, o scores determinados pelas combinações lineares podem ser calculados para cada elemento amostral.

2.4 ANÁLISE DE AGRUPAMENTOS

A finalidade da análise de agrupamentos, também conhecida como análise de conglomerados, classificação ou cluster é dividir os elementos de um conjunto de dados em grupos similares entre si e heterogêneos entre os grupos diferentes, levando em conta as características medidas [12].

2.4.1 MEDIDAS DE DISSIMILARIDADE

As medidas de dissimilaridade são uma etapa importante na análise de agrupamentos. Diversos tipos de medidas podem ser utilizadas para mensurar a dissimilaridade entre dois elementos, como a distância Euclidiana, distância Euclidiana ao quadrado, distância generalizada

e distância de Mahalanobis [12].

Nesse trabalho, a realização da análise de agrupamentos será utilizada a distância euclidiana. Esse tipo de distância é baseada no princípio de que dois pontos são similares quanto menor for a distância entre eles.

A distância euclidiana quadrática entre dois elementos X_l e X_k com $l \neq k$ é representada por :

$$d(X_l, X_k) = \sqrt{\sum_{i=1}^p (X_{il} - X_{ik})^2} \quad (2.13)$$

2.4.2 AGRUPAMENTOS NÃO HIERÁRQUICOS E MÉTODO K-MEANS

Os métodos não hierárquicos buscam dividir um conjunto de n elementos em k grupos (clusters) de forma direta, levando em consideração a semelhança entre os elementos de um mesmo grupo e a diferença entre os grupos. Para obter a melhor divisão em k grupos, é necessário utilizar algum tipo de critério para avaliar a qualidade da divisão [12].

Os métodos não hierárquicos requerem que o número de clusters seja especificado previamente. Durante o processo de agrupamento, os novos grupos são formados através da divisão ou união de grupos já existentes. Esses métodos são iterativos e possuem capacidade para lidar com grandes conjuntos de dados. Os algoritmos mais conhecidos são o K-means e o Fuzzy c-Means.

O método K-Means, desenvolvido por Hartigan e Wong (1979), é amplamente utilizado e conhecido em problemas reais. Ele funciona alocando cada elemento amostral ao cluster cujo vetor de média amostral é o mais próximo do vetor de valores observados para o respectivo elemento. O algoritmo é composto por quatro passos representados a seguir.

- **Passo 1 :** Para iniciar o processo de divisão, é necessário escolher k pontos iniciais, chamados de sementes, que serão utilizados como centróides.
- **Passo 2 :** Depois de escolher as k sementes, cada elemento do conjunto de dados é comparado com cada centróide inicial utilizando uma medida de distância, como a distância Euclidiana. Cada elemento é então atribuído ao grupo cuja distância é a menor.
- **Passo 3 :** Em seguida, para cada um dos n elementos amostrais, recalcula-se os valores dos centróides para cada novo grupo formado, e repete-se o passo 2, considerando os centróides destes novos grupos.
- **Passo 4 :** Os passos de atribuição e atualização devem ser repetidos até que todos os elementos amostrais estejam bem distribuídos nos grupos, ou seja, nenhuma mudança adicional seja necessária para alocar os elementos nos grupos corretos.

2.4.3 CRITÉRIOS PARA SELEÇÃO DO NÚMERO DE CLUSTERS

Os critérios de seleção de grupos nos métodos não hierárquicos são importantes porque ajudam a avaliar a qualidade dos clusters gerados pelo algoritmo. Esses critérios ajudam a determinar se os elementos estão bem agrupados e se os grupos são distintos e coerentes. O critério mais utilizado para a escolha do número de agrupamentos no algoritmo k-means é o método Elbow.

O método de Elbow ("cotovelo") é uma técnica utilizada para determinar o número ideal de grupos a serem formados em uma análise de agrupamento. Ele funciona plotando o valor da soma dos quadrados dentro dos grupos versus o número de grupos e identificando o ponto onde a redução da soma de quadrados diminui significativamente. Esse ponto é considerado o número ideal de grupos a serem formados [8].

Além do critério para a tomada de decisão do número de grupos, existem métodos que indicam a qualidade dos agrupamentos gerado pelo algoritmo. O quociente entre a soma dos quadrados entre os grupos e a soma total dos quadrados, conhecido como R-quadrado, é uma das maneiras de avaliar a qualidade dos clusters. A soma dos quadrados entre grupos mede a variação entre os grupos, assim quanto maior o valor, maior será a diferença entre os agrupamentos. Já a soma total dos quadrados mede a variação total dos dados, assim quanto menor o valor, maior qualidade nos grupos formados. Desse modo, o resultado do R-quadrado será em porcentagem e um valor acima de 0,7 indica melhor qualidade dos grupos formados. [8]

3. RESULTADOS

3.1 ANÁLISE DESCRITIVA DOS FUNDOS IMOBILIÁRIOS

A princípio foi realizado uma análise sobre as características dos fundos quanto aos setores presentes no mercado. De acordo com a Figura 3.1, o setor com maior participação dentre os ativos analisados é o de títulos e valores imobiliários (43%) em que são fundos que aportam recursos em títulos como LCI, CRI ou ações de empresas [9]. A menor participação são dos fundo de hospitais (1%).

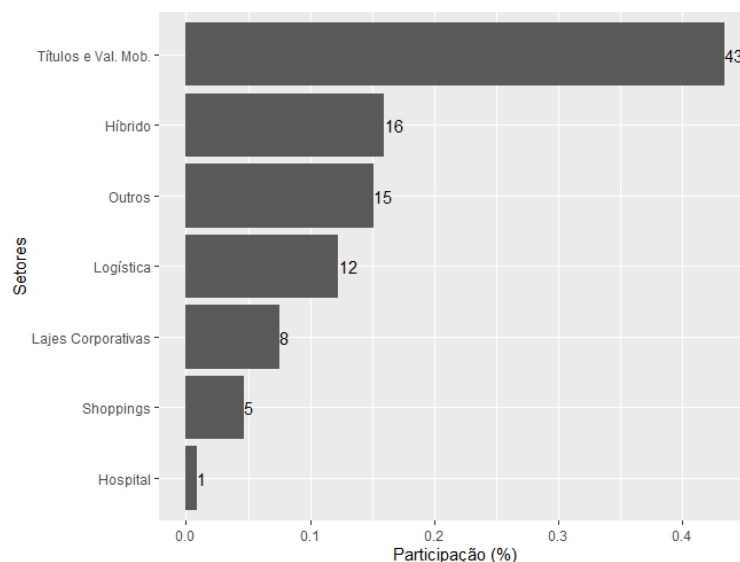


Figura 3.1: Gráfico de barras da relação dos fundos imobiliários por setor

A Tabela 3.1 apresenta estatísticas descritivas das variáveis de estudo, como média, desvio padrão, coeficiente de variação, quartil 1, mediana, quartil 3, mínimo e máximo para melhor compreensão da distribuição dos dados. Num contexto geral, é possível observar que os índices são heterogêneos, pois a maioria apresentou coeficiente de variação acima de 0,40. Alguns chegaram a ter valores de CV próximo de 1, mostrando a grande variedade de perfis presentes no banco de dados. A exceção é o índice preço sobre valor patrimonial por cota (VPC) tem uma distribuição homogênea, com CV de 0,09.

No geral, as médias se aproximam mais do mínimo das variáveis, indicando uma distribuição assimétrica. Portanto podemos dizer que o maior volume dos fundos estão com valores de seus índices abaixo da média, existindo alguns índices com valores muito acima, destoando da

maioria.

Os máximos das variáveis preço e VPC estão muito acima da média. Além disso, também indica que será necessário padronizar os índices financeiros para realizar as análises.

Tabela 3.1: Tabela com estatísticas descritivas índices financeiros

Variáveis	Média	Desvio Padrão	CV	Q1	Mediana	Q3	Mínimo	Máximo
Preço	69,82	34,47	0,49	60,82	77,88	85,32	5,67	286,16
VPC	95,75	45,35	0,47	94,08	99,31	105,03	8,96	351,18
Patrimônio Líquido	629.186.551	551.303.646	0,88	238.032.236	484.262.541	953.154.960	45.450.869	3.392.484.754
Dividendos	0,79	0,47	0,6	0,59	0,73	0,95	0,06	3,19
P/VPC	0,99	0,09	0,09	0,94	0,99	1,03	0,79	1,33
Retorno	0,05	0,09	1,66	0,03	0,04	0,05	0,01	0,77
Volume	33.517.729	26.933.975	0,8	12.416.842	26.537.495	48.574.010	3.046.523	123.671.066
Risco	0,16	0,06	0,36	0,12	0,16	0,19	0,05	0,46

3.2 ANÁLISE DAS COMPONENTES PRINCIPAIS

Neste trabalho, será apresentado os resultados da utilização da técnica de análise de componentes principais (PCA) em um conjunto de dados de índices financeiros de fundos imobiliários. O objetivo é compreender as principais características e tendências presentes nesses dados.

A Tabela 3.2 mostra os autovalores gerados pela matriz de correlação, a porcentagem da variação explicada por cada componente principal (CP) e a porcentagem acumulada da variação explicada. Esses valores são obtidos através da adição sucessiva das proporções de variação explicada de cada CP, até atingir o valor total de 100%.

Os resultados mostram que cerca de 35,54% da variação total é explicada pela primeira componente principal. Além disso, 84,79% da variação nos dados é explicada pelas 4 primeiras componentes. De acordo com o critério de Kaiser, as 4 primeiras componentes com autovalores maiores que 1 serão mantidas. Isso significa que a redução de 8 variáveis originais para 4 variáveis criadas a partir das combinações lineares leva à perda de cerca de 15,22% da informação na variação dos dados, o que é considerado uma pequena perda.

Tabela 3.2: Tabela dos autovalores com percentuais da variação para cada componente

Componentes Principais	Autovalores	% Variância	% Variância Acumulada
CP1	2,843	35,540	35,540
CP2	1,803	22,539	58,078
CP3	1,121	14,007	72,085
CP4	1,016	12,699	84,784
CP5	0,592	7,402	92,186
CP6	0,346	4,326	96,511
CP7	0,197	2,458	98,969
CP8	0,082	1,031	100,000

A partir das primeiras 4 componentes principais, é possível extrair os autovetores criados a partir das combinações lineares, assim compreendendo a importância relativa de cada variável nas componentes. A Tabela 3.3 apresenta os valores dos coeficientes. Os valores vazios significam que o coeficiente para determinada variável está abaixo de 0,1.

Tabela 3.3: Tabela dos coeficientes para cada componentes e variáveis

	CP1	CP2	CP3	CP4
Preço	0,517	0,105	-	0,187
VPC	0,519	0,203	0,183	0,123
Patrimônio Líquido	0,220	-0,596	0,297	-0,128
Dividendos	0,443	0,302	-	-
P/VPC	0,406	-	-0,440	-0,261
Retorno	-	0,230	-	-0,910
Volume	0,225	-0,613	0,165	-0,170
Risco	-	0,269	0,810	-

- **Componente Principal 1 :**

$$Y_1 = 0,517(\text{Preço}) + 0,519(\text{VPC}) + 0,220(\text{PL}) + 0,443(\text{Dividendos}) + 0,406(\text{P/VPC}) + 0,225(\text{Volume}) \quad (3.1)$$

É possível observar na primeira componente principal que os índices preço, VPC, P/VPC e dividendos tem maiores contribuições para a variação explicada pela primeira componente. Além disso, as variáveis preço, VPC, patrimônio líquido, dividendos, P/VPC e volume estão altamente relacionadas entre si. Isso pode indicar, por exemplo, que ativos com valores elevados dessas variáveis tendem a seguir uma tendência semelhante. Também pode ser interpretado como a tendência de que esses fundos imobiliários são atrativos para investidores, que estão dispostos a pagar um preço alto pelo ativo, tem um patrimônio líquido e dividendo elevado e são mais negociadas na bolsa.

Em geral, o valor patrimonial por cota (VPC) é calculado dividindo o patrimônio líquido do fundo imobiliário pelo número de cotas em circulação. Como o VPC reflete a situação financeira do fundo, ele é usado como uma medida de avaliação da qualidade do ativo. Quando o VPC é alto, isso significa que o patrimônio líquido do fundo é alto em relação ao número de cotas em circulação, o que geralmente é visto como positivo e pode aumentar a procura pelas cotas, o que por sua vez pode aumentar o preço das cotas, os dividendos e o P/VPC.

- **Componente Principal 2 :**

$$Y_2 = 0,105(\text{Preço}) + 0,203(\text{VPC}) - 0,596(\text{PL}) + 0,302(\text{Dividendos}) + 0,230(\text{Retorno}) - 0,613(\text{Volume}) + 0,269(\text{Risco})$$

A segunda componente principal tem coeficientes que indicam uma relação negativa com maior contribuição do patrimônio líquido e volume, e relações positivas entre VPC, dividendos, retorno e risco. Um score alto nessa componente sugere que os valores dos fundos imobiliários estão relacionados com valores de patrimônio líquido e volume baixo. Um score baixo sugere o contrário, ou seja, os valores dos fundos imobiliários estão relacionados com patrimônio líquido e volume de alto valores.

Além disso, faz sentido que as variáveis com coeficientes negativos na componente 2 estejam relacionadas pois, no mercado de fundos imobiliários, se o patrimônio líquido de um fundo aumenta, isso pode levar a um aumento no volume de negociações, pois os investidores podem considerar o fundo mais atrativo. Por outro lado, se o patrimônio líquido diminui, isso pode resultar em uma diminuição nas transações.

- **Componente Principal 3 :**

$$Y_3 = 0,183(VPC) + 0,297(PL) - 0,440(P/VPC) + 0,165(Volume) + 0,810(Risco) \quad (3.2)$$

A terceira componente apresentou uma relação inversa entre risco e preço pelo valor patrimonial por cota (P/VPC). De acordo com os resultados, o risco foi a variável com a maior contribuição positiva nessa componente, enquanto o P/VPC teve a segunda maior contribuição negativa. Isso sugere que, no geral, os fundos imobiliários com maior risco tendem a ter uma relação P/VPC menor, ou seja, um preço menor em relação ao seu valor patrimonial por cota.

Em geral, faz sentido o P/VPC estar relacionado inversamente com o risco. O P/VPC é uma medida de valor, que indica quanto um investidor está disposto a pagar pelo ativo em relação ao lucro gerado. Quanto maior o risco, menor é o retorno, o que pode levar a uma diminuição no P/VPC. Por outro lado, quando o risco é baixo, o rendimento é maior, o que pode levar a um aumento no P/VPC. Então, é esperado que o P/VPC seja inversamente proporcional ao risco.

- **Componente Principal 4 :**

$$Y_4 = 0,187(Preco) + 0,123(VPC) - 0,128(PL) - 0,261(P/VPC) - 0,910(Retorno) - 0,170(Volume)$$

A quarta componente está relacionada ao desempenho de retorno dos fundos imobiliários. O coeficiente negativo de $-0,910$ para retorno indica que quanto mais alto o retorno, menor será o score dessa componente. Já os coeficientes positivos para preço e VPC e negativos para PL, P/VPC e volume indicam que essas variáveis também estão relacionadas com o desempenho de retorno, mas de maneira menos direta do que o retorno em si.

Em geral, pode-se interpretar essa componente como uma medida geral de desempenho de retorno dos fundos imobiliários, levando em conta tanto o retorno quanto outras variáveis que podem afetar esse desempenho.

3.3 ALGORITMO K-MEANS

As quatro componentes principais foram selecionadas e compreendidas na seção anterior. Agora, os scores dessas componentes serão utilizados para realizar o agrupamento das variáveis dos fundos imobiliários utilizando o algoritmo de k-means. Esse algoritmo irá agrupar os

dados de acordo com as similaridades entre as componentes, permitindo uma melhor análise e compreensão do conjunto de dados.

Após a realização da análise dos dados, foi escolhido 7 clusters. A tomada de decisão para a escolha desse número de grupos foi realizada utilizando o método Elbow e o R-quadrado. O método Elbow, representado pela Figura 3.2, é possível verificar que o aumento da soma de quadrados entre clusters começa a diminuir após o número de agrupamentos igual 7, indicando que adicionar mais clusters não trará benefícios para a análise.

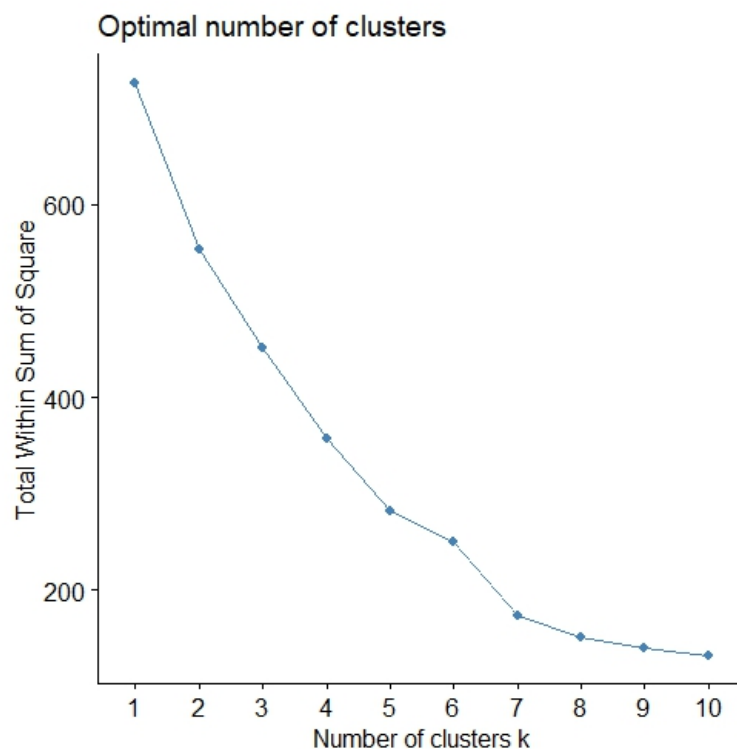


Figura 3.2: Gráfico do cotovelo para seleção de clusters.

A Tabela 3.4 apresenta a soma dos quadrados entre clusters, juntamente com o R-quadrado, quando o algoritmo K-Means foi aplicado para grupos de 4 a 10. O R-quadrado é uma medida de quão bem os dados se ajustam aos clusters, quanto maior o R-quadrado, melhor é a capacidade do modelo para explicar a variação dos dados. Pode-se observar que o R-quadrado aumenta à medida que o número de clusters aumenta até 7, alcançando seu pico com 77.2%. Após isso, o R-quadrado se mantém constante.

Tabela 3.4: Tabela Soma de Quadrados entre Cluster e R-quadrado de 4 a 10 grupos

Número de grupos	SSB	R^2
4	381,472	0,526
5	458,827	0,632
6	526,961	0,726
7	560,314	0,772
8	585,907	0,807
9	597,500	0,823
10	609,435	0,840

Portanto, a combinação desses dois métodos, Elbow e R^2 permitiu chegar ao número ótimo de clusters igual a 7 para a melhor análise e compreensão dos fundos imobiliários.

A Tabela 3.5 demonstra a frequência dos ativos em cada cluster. Observa-se que o grupo com maior número de fundos é o cluster 2 seguido pelo cluster 6. Os clusters 3, 5 e 7 apresentaram-se ativos que escapam do padrão, possivelmente outliers.

Tabela 3.5: Tabela de Frequência de fundos imobiliários em cada cluster

Clusters	Freqüência
Cluster 1	13
Cluster 2	34
Cluster 3	1
Cluster 4	23
Cluster 5	2
Cluster 6	33
Cluster 7	1

A análise dos clusters formados com poucos fundos imobiliários pode fornecer informações valiosas sobre as características dos diferentes tipos de ativos. A Tabela 3.6 mostra os ativos para cada cluster inserido e as respectivas médias dos scores para cada componente principal.

Ao analisar os scores do cluster 3, é possível observar que ele tem pontuação alta nas componentes 2 e 3. Isso indica que esses ativos são fora do comum, pois possuem baixo volume de negócios e patrimônio líquido (CP2) e risco elevado (CP3). Já o cluster 5 tem como principal característica a pontuação alta na componente 4, indicando que são fundos com alto retorno (CP4), mas baixo volume negociado na bolsa e patrimônio líquido (CP3). O cluster 7 é um outlier no aspecto de preço e valor patrimonial por cota, indicando um fundo caro.

Tabela 3.6: Tabela dos clusters com baixa frequência de fundos

Cluster	Ativo	CP1	CP2	CP3	CP4
3	CARE11	0,927	4,448	4,600	-0,831
5	RCRB11	1,258	2,781	-0,562	-4,910
5	VRTA11	0,424	2,517	-1,303	-7,785
7	FIB11	8,658	3,801	-0,612	1,625

Em seguida, será realizado a interpretação dos cluster com maior quantidade de fundos imobiliários, representados pela Tabela 3.7 com as médias dos scores para componentes.

Tabela 3.7: Tabela dos grupos com alta frequência de fundos

Cluster	Quantidade de ativos	CP1	CP2	CP3	CP4
1	13	-3,005	-0,458	-0,624	-0,324
2	34	0,621	0,400	-0,774	0,353
4	23	1,271	-1,743	0,438	-0,287
6	33	-0,683	0,573	0,673	0,325

- **Grupo 1 :** A principal característica desse cluster é a média do score da componente 1, indicando que são fundos com valores menores em preço, valor patrimonial por cota (VPC), dividendos e P/VPC. Além disso, as componentes 2, 3 e 4 apresentam valores de volume e patrimônio líquido, risco e retorno moderadamente altos, sugerindo que esses fundos são considerados mais baratos no mercado. O cluster 1 pode ser recomendado para investidores que optam por aportar ativos de baixo preço.
- **Grupo 2 :** Este cluster apresenta fundos com características moderadas em preço, volume, patrimônio líquido, risco e retorno. Não há uma característica relevante que os defina. No entanto, comparado ao outros grupos da Tabela 3.7, são fundos com maior média de score absoluta para as componentes 3 e 4. Além disso, esses fundos são adequados para investidores que buscam um equilíbrio moderado entre risco e retorno.
- **Grupo 4 :** Ativos com preços, dividendos, P/VPC e VPC relativamente altos. Além disso, possuem perfil de baixo risco e retorno moderado, e apresentam volumes de negociações e patrimônio líquido elevados em comparação com os outros clusters. Além do mais, esses fundos são atrativos para investidores que estão dispostos a pagar um preço moderadamente mais alto, com um equilíbrio entre risco e retorno razoavelmente bom.
- **Grupo 6 :** Fundos neste cluster apresentam valores de preço e dividendos abaixo da média do mercado. Eles também possuem um perfil de retorno menores e riscos elevados quando comparado aos outros agrupamentos. O cluster 6 apresenta características desfavoráveis para investidores, pois a relação entre risco e retorno é desfavorável.

4. CONCLUSÕES

A análise de componentes principais (PCA) e o agrupamento K-means foram dois métodos multivariados utilizados para realizar a análise exploratória dos dados relacionados ao mercado de fundos imobiliários. A PCA permitiu identificar relações entre as variáveis originais, permitindo a redução de oito variáveis para quatro componentes principais que combinam as variáveis originais de tal forma que sejam o mais independentes entre si e que explicam a maior variação dos dados. Dessa forma, reduzindo a complexidade de interpretação quando analisado com os dados originais.

Já o agrupamento K-means foi utilizado para identificar sete grupos de fundos imobiliários com características semelhantes dentro de cada cluster, mas com perfis diferentes para cada agrupamento. Além disso, foi possível verificar quatro fundos que fogem do padrão de mercado. Durante o processo de análise de cada cluster, foi possível relacionar os grupos com perfis de investidores diferentes.

Por fim, os dois métodos multivariados mostraram-se eficientes ao realizar a análise exploratória dos dados relacionados ao mercado de fundos imobiliários, tornando-se mais uma ferramenta importante que auxilia investidores no processo de seleção de melhores investimentos. Esses métodos possibilitam a identificação de padrões e relações entre as variáveis, assim como o agrupamento de dados similares, o que é essencial para a tomada de decisão do investidor.

REFERÊNCIAS BIBLIOGRÁFICAS

- [1] *Bolsa de Valores*. <https://www.dicionariofinanceiro.com/bolsa-de-valores/>, acessado em 08/11/2022.
- [2] *Economática*. <https://economatica.com/saddin>, acessado em 08/01/2023.
- [3] *Mercado Financeiro*. <https://www.dicionariofinanceiro.com/mercado-financeiro/>, acessado em 08/11/2022.
- [4] *The R Project for Statistical Computing*. <https://www.r-project.org/>, acessado em 08/01/2023.
- [5] *Yahoo Finance*. <https://finance.yahoo.com/>, acessado em 08/01/2023.
- [6] *História do mercado de capitais*, 2015. <https://www.gov.br/investidor/pt-br/investir/como-investir/conheca-o-mercado-de-capitais/historia-do-mercado-de-capitais>, acessado em 08/11/2022.
- [7] *Raio-X do investidor*, 2021. https://www.anbima.com.br/pt_br/especial/raio-x-do-investidor-2021.html, acessado em 08/11/2022.
- [8] Aggarwal, C. C. K., .R. C.: *Data Clustering: Algorithms and Applications*. Chapman & Hall, p. 616, 2013.
- [9] Danilos, B. M..B.: *Guia Fundos Imobiliários : Introdução sobre investimentos seguros e rentáveis*, vol. 1. Vivalendo, 1ª ed., 2018.
- [10] Eric, D.: *Aprenda a Investir em Fundo Imobiliários*, vol. 1. Copyright, 1ª ed., 2019.
- [11] Reis, T.: *Tudo sobre Bolsa de Valores*. <https://www.suno.com.br/guias/bolsa-de-valores/>, acessado em 08/11/2022.
- [12] Sueli, A. M.: *Análise de Dados Através de Métodos de Estatística Multivariada*, vol. 1. Editora UFMG, 1ª ed., 2005.
- longtable

A. APÊNDICE

Tabela A.1: Tabela de fundo imobiliários coletados

Código FIIS	Setor
AFHI11	Títulos e Val. Mob.
AIEC11	Lajes Corporativas
ALZR11	Logística
ARCT11	Híbrido
ARRI11	Títulos e Val. Mob.
BARI11	Títulos e Val. Mob.
BBPO11	Lajes Corporativas
BCFF11	Títulos e Val. Mob.
BCIA11	Títulos e Val. Mob.
BCRI11	Títulos e Val. Mob.
BLMG11	Logística
BLMR11	Títulos e Val. Mob.
BFFF11	Títulos e Val. Mob.
BRCO11	Logística
BRCR11	Híbrido
BTAL11	Outros
BTLG11	Logística
BTRA11	Outros
CACR11	Híbrido
CPFF11	Híbrido
CPTS11	Títulos e Val. Mob.
CVBI11	Títulos e Val. Mob.
DEVA11	Títulos e Val. Mob.
GTWR11	Lajes Corporativas
HABT11	Títulos e Val. Mob.
HCTR11	Outros
HFOF11	Títulos e Val. Mob.
HGBS11	Shoppings

Continuação da Tabela

Código FIIS	Setor
HGCR11	Títulos e Val. Mob.
HGFF11	Títulos e Val. Mob.
HGLG11	Logística
HGRE11	Lajes Corporativas
HGRU11	Híbrido
HSAF11	Títulos e Val. Mob.
HSLG11	Logística
HSML11	Shoppings
IRDM11	Títulos e Val. Mob.
JSRE11	Híbrido
KFOF11	Títulos e Val. Mob.
KISU11	Títulos e Val. Mob.
KNCR11	Títulos e Val. Mob.
KNHY11	Títulos e Val. Mob.
KNIP11	Títulos e Val. Mob.
KNRI11	Híbrido
KNSC11	Títulos e Val. Mob.
LVBI11	Logística
MALL11	Shoppings
MCHF11	Títulos e Val. Mob.
MFII11	Híbrido
MGFF11	Títulos e Val. Mob.
MORE11	Títulos e Val. Mob.
MXRF11	Híbrido
NSLU11	Hospital
OUJP11	Títulos e Val. Mob.
PATL11	Híbrido
PLCR11	Híbrido
QAGR11	Outros
RBFF11	Títulos e Val. Mob.
RBRF11	Títulos e Val. Mob.
RBRL11	Logística
RBRP11	Outros
RBRR11	Títulos e Val. Mob.
RBRY11	Títulos e Val. Mob.
RBVA11	Outros
RCRB11	Lajes Corporativas
RECR11	Títulos e Val. Mob.
RECT11	Híbrido

Continuação da Tabela

Código FIIS	Setor
RVBI11	Títulos e Val. Mob.
SADI11	Títulos e Val. Mob.
SARE11	Híbrido
SDIL11	Logística
SNCI11	Títulos e Val. Mob.
SNFF11	Títulos e Val. Mob.
SPTW11	Lajes Corporativas
TGAR11	Outros
TORD11	Outros
URPR11	Outros
VCJR11	Títulos e Val. Mob.
VGHF11	Títulos e Val. Mob.
VGIP11	Outros
VIFI11	Títulos e Val. Mob.
VILG11	Logística
VINO11	Lajes Corporativas
VISC11	Shoppings
VIUR11	Títulos e Val. Mob.
VSLH11	Outros
VTLT11	Logística
XPCI11	Outros
XPIN11	Outros
XPLG11	Logística
XPML11	Shoppings
XPPR11	Outros
XPSF11	Outros
