INSTITUTO UNIVERSITÁRIO DE LISBOA

Curso de Licenciatura em Ciência de Dados

2º ano/ 1º semestre - 2024/2025

Unidade Curricular:

Métodos de Aprendizagem Não Supervisionada

Segmentação de Casas

Grupo 4:

António Santos, nº 123434;

Frederico Silva, nº 112959

Gonçalo Henriques, nº 123422;

José Alberto nº 121959;

Jonasse Mbaki, nº 111900;

Docente: Mafalda Ponte

Lisboa, dezembro de 2024

ÍNDICE

1. INTRODUÇÃO

No âmbito da unidade curricular Métodos de Aprendizagem não Supervisionada, durante o primeiro semestre 2024/2025, foi proposta a realização de um relatório referente a um projeto que consiste na análise de uma base de dados em que as variáveis dizem respeito a características de fogos, as quais irão ser mencionadas mais à frente. Dado que esta análise é realizada num contexto não supervisionado e visando uma tomada de decisão informada, o presente trabalho tem como objetivo a segmentação das casas em grupos de características semelhantes e, a análise do impacto da remodelação/ano de construção da mesma sobre o seu preço final dentro do mercado imobiliário.

A importância deste estudo reside na possibilidade de compreender os fatores que causam variações no preço das propriedades, seja através do aumento ou da diminuição do seu valor. A compreensão destes fatores é essencial para diversos participantes do mercado imobiliário, incluindo compradores, vendedores, investidores e agentes imobiliários, uma vez que capacita estes agentes a tomar decisões mais estratégicas e informadas, otimizando os seus retornos neste mercado imobiliário.

Com intuito de alcançar os objetivos mencionados, iniciou-se uma análise exploratória detalhada das variáveis, de modo a tratar dos valores que exigiram o seu respetivo tratamento e a identificar possíveis outliers que poderiam afetar os resultados. Posteriormente foi aplicada uma análise PCA (Principal Componente Analysis), uma técnica estatística que permite reduzir a dimensionalidade dos dados, preservando a maior quantidade de informação possível. Através de uma análise dos scores dos componentes principais, PC (Principal Components), foi possível identificar agrupamentos (Clusters) e padrões dentro dos dados, isto por si, facilitando a compreensão das características que mais predominam no mercado do setor imobiliário e que são mais desejadas.

2. DADOS

Do dataset fornecido existia um total de 4.600 registos e 49 variáveis, contudo a partir da variável "Ano_renovacao", as subsequentes eram binárias, destinadas a selecionar a base de dados final a utilizar para cada grupo. Deste modo, como fazemos parte do grupo "T2Gr04", após a seleção da respetiva variável, totalizou-se, numa primeira fase, 4147 registos. A caracterização das variáveis em estudo, ativas e passivas, encontra-se na subsecção [2.2].

2.1. Tratamento das variáveis

O tratamento das variáveis foi realizado, com o auxílio de visualizações gráficas (boxplots e histogramas, encontrados nos anexos 1 e 2) das variáveis individualmente, de modo a garantir consistência e representatividade dos dados, para uma análise mais confiável. Apenas não exigiram o tratamento das variáveis: "nrQuartos", "vista" e "condicao". Em relação ao "nrWC" abordou-se do seguinte modo: dado que as incrementações dos valores são de 0.25, considerou-se que para cada uma das mesmas, após o valor inteiro, referente ao número de casas de banho completas, corresponderia a uma "WC" que por definição "Water Closet" refere-se a um espaço mais restrito (contém apenas a sanita e por vezes o lavatório). Por exemplo, se o valor for 1.75 corresponde no total a 4 espaços sanitários. De modo paralelo para a variável "nrAndares", dado que as incrementações são de 0.5, considerou-se que após o valor inteiro, referente ao número de andares em si, corresponderia ao sótão. Referente ao "Ano_construcao" e "Ano_renovacao" nos casos em que estes não eram sequentes foi atribuído o valor "0" para o ano de renovação.

Para tornar as análises ainda mais precisas foi crucial a remoção dos valores que se considerou como outliers, valores estes que se encontravam para além dos limites máximos definidos para cada variável relevante. O preço ficou limitado a menos de 2.5 milhões de unidades monetárias, a área da sala de estar a menos de 180 m², o tamanho do lote a menos de 20.000 m², a área do piso a menos de 500 m² e, a área da arrecadação a menos de 750 m². Este passo foi fundamental pois permitiu a eliminação de dados que poderiam causar distorções nos resultados das análises posteriores realizadas. A base de dados final utilizada contém assim um total de 4.099 observações após a exclusão de 48 registos considerados outliers.

2.2. Variáveis de INPUT e PROFILE

Entre as variáveis disponíveis na base de dados final, as de perfil, usadas para caracterizar os clusters formados durante a análise, são indicadores cruciais no estudo do impacto das características das residências sobre o preço final. Estas incluem o Preço da Residência (Preco), o Ano de Construção (Ano_construcao) e o Ano de Renovação/Remodelação da Residência (Ano_renovacao), essenciais para analisar o impacto de fatores temporais e reformas no valor das residências. Já as variáveis ativas (INPUT), utilizadas para segmentação, abrangem características sobre dimensões das propriedades, como número de quartos (nrQuartos), número de espaços sanitários (nrWC), número de andares totais (nrAndares), área da sala de estar (Sala_estar_m2), área total da residência (lote_m2), vista que a residência apresenta (vista), a condição da mesma (condição), área no interior da residência incluindo todos os andares (Piso_m2) e a área da arrecadação (arrecadação_m2).

3. IDENTIFICAÇÃO DAS DIMENSÕES DA ANÁLISE

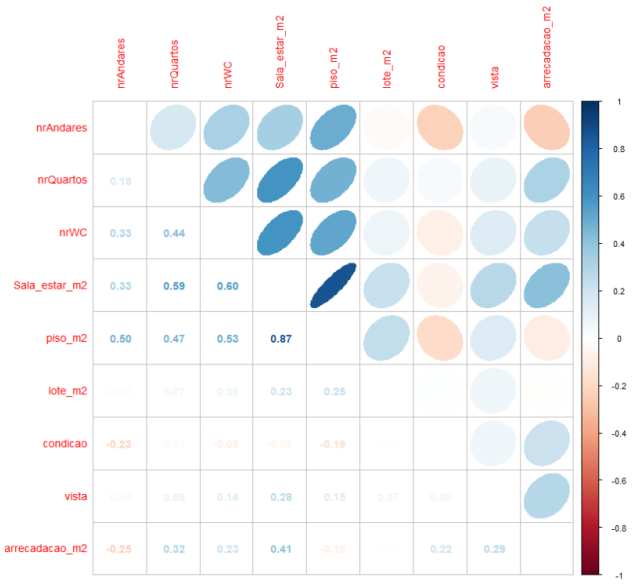
Para este passo será utilizado o PCA (Principal Components Analysis). O mesmo segue um conjunto de passos que devem ser seguidos de modo a garantir que a mesma é uma opção viável para o nosso dataset em estudo. Estes passos serão descritos de seguida.

3.1. Seleção das variáveis de input

Foram selecionadas um total de 9 variáveis, as mesmas que foram mencionadas como as variáveis de Input, assumindo que seguem uma distribuição aproximadamente Normal. Estas serão as variáveis que utilizaremos para a aplicação do PCA, ficando de parte as variáveis que dizem respeito aos preços das casas, ano de construção e ano de renovação.

3.2. Adequabilidade (avaliar a aplicabilidade do PCA):

Matriz de correlação: Observou-se que existem valores que se encontram dentro do intervalo onde não é adequado aplicar o PCA, [-0.3; 0.3] mas também se verificou que há casos onde se encontram fora deste intervalo pelo que é possível considerar que pode haver a possibilidade de as variáveis serem correlacionadas (condição necessária para aplicação do PCA), contudo a matriz não é suficiente para determinar a adequabilidade do PCA ao dataset em estudo.



Mapa de correlações

Bartlett's Test of Sphericity: Observou-se que o valor de p-value é bastante inferior ao valor de referência do nível de significância (0.05) pelo que se rejeita a

hipótese de as variáveis serem não correlacionadas (hipótese nula). Deste modo, por este critério a aplicabilidade do PCA poderá ser adequada ao dataset em estudo.

Medida KMO: A medida de "Overall MSA" encontra-se entre os valores considerados inaceitáveis e medíocres, pelo que se considerou um valor medíocre, o que não descarta a adequabilidade do PCA para a amostra em estudo.

```
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = correlation)
Overall MSA =  0.53
MSA for each item =
     nrQuartos           nrWC  Sala_estar_m2
          0.96           0.95           0.48
       piso_m2 arrecadacao_m2
          0.45           0.21
```

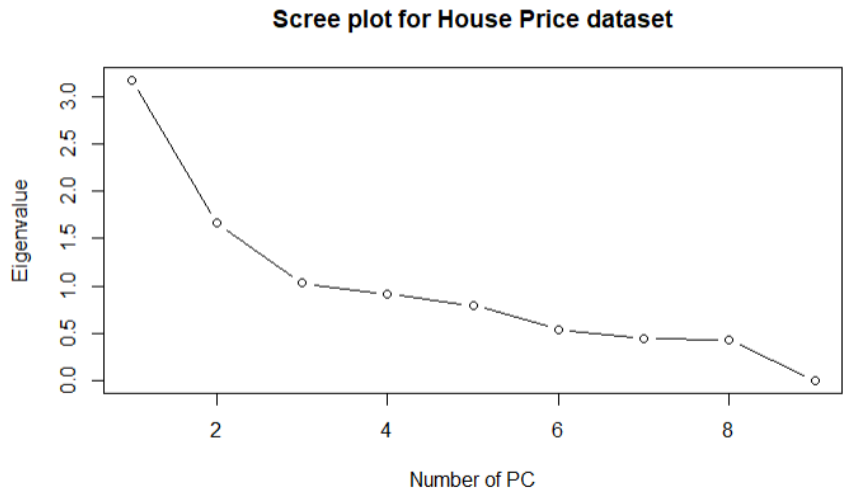
Tabela parcial dos resultados do KMO (tabela completa no anexo 3)

3.3. Número de componentes e extração dos PC's

Conhecimento teórico: Dado que não possuímos à priori nenhum conhecimento sobre o mercado imobiliário de modo a definir a quantidade de PC a extrair do dataset, este passo não será tomado em conta. Mas, para os passos seguintes, os mesmos foram aplicados com todas variáveis de Input que se definiram de modo a, por este caminho ser possível, de acordo com critérios a seguir definir o número de componentes ideal.

Os passos que se seguem serão feitos com as variáveis estandardizadas isto porque a aplicação do PCA torna-se mais eficaz quando os dados estão centrados e a estandardização é uma forma de centrar os dados. O número de componentes principais a serem considerados inicialmente serão 9 pois é o número de variáveis de input, mencionadas anteriormente no ponto em que se faz referência a Seleção das variáveis de input.

Critério de Kaiser: Este critério consiste na ideia de se selecionar como componentes principais aqueles que têm os seus valores próprios superiores a 1. Pode-se observar este fenómeno de dois modos, por intermédio de um gráfico e por tabela (para garantir maior precisão).




```
[1] 3.177 1.663 1.030 0.918 0.796 0.542 0.447 0.427 0.000
```

Tabela de valores próprios por componentes ordenados

Por este critério a melhor aposta seria escolher-se 3 componentes principais. De modo a garantir precisão, vai ser testado também os demais critérios a ver se esta continua a ser a decisão mais acertada a se tomar.

Critério da variância explicada: Este critério consiste em selecionar o número de componentes de acordo com a variância acumulada até no mínimo 60%, ou seja, seleciona-se o número de componentes até ao qual é possível explicar até pelo menos 60% dos dados que se tem na nossa amostra. Pode-se observar os resultados apresentados na tabela a seguir e tirar conclusões:

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
SS loadings	3.177	1.663	1.030	0.918	0.796	0.542	0.447	0.427	0
Proportion Var	0.353	0.185	0.114	0.102	0.088	0.060	0.050	0.047	0
Cumulative Var	0.353	0.538	0.652	0.754	0.843	0.903	0.953	1.000	1

Tabela de loadings e variâncias dos componentes principais

A seguinte avaliação será baseada apenas na terceira linha da tabela acima apresentada, a qual representa o valor da variância cumulativo. É possível observar que se pode explicar pelo menos 60% dos dados na amostra com precisamente 3 componentes principais o que vai de acordo ao observado no critério anterior. Dado que os dois critérios decidiram pelo mesmo número de componentes não foi necessário avançar para os próximos critérios com o intuito de "desempate". Podemos então observar apenas o comportamento das variáveis com o número de componentes indicados pelos critérios anteriores por meio das suas comunalidades.

Critério das comunalidades (Apenas para observar o comportamento das variáveis): Este critério consiste em avaliar de acordo ao número de componentes definidos quais comunalidades das variáveis de input. Na sua interpretação, define-se como o número de componentes ideal, aquele em que as variáveis são todas iguais ou superiores a 50%. Isto é possível ser visto e interpretado em tabela, tabela esta que dado o número de colunas estará disponível apenas parcialmente logo a seguir e completa no anexo 4 e no script enviado.

nrAndares	vista	condicao
0.6300000	0.2581294	0.3655662

Tabela parcial das comunalidades

A tabela de comunalidades revela que pelo menos duas variáveis, "vista" e "condição", possuem valores inferiores a 50%. Isso é consistente com o fato de que essas variáveis não apresentam uma correlação significativa com as demais, conforme observado na matriz de correlações. Embora esta análise não seja determinante para a escolha final do número de componentes principais, ela indica que estas variáveis não se ajustam bem ao modelo de PCA. Considerando isso, decidimos prosseguir com o mesmo número de componentes e avançar para o próximo passo.

3.4. Interpretação e Rotação dos Componentes

A análise seguiu diretamente para a avaliação dos loadings, uma vez que não houve necessidade de reconsiderar ou ajustar o número de Componentes a ser extraído dado que os critérios anteriores realizados relataram a necessidade do mesmo número de componentes (3). Para a interpretação dos mesmos, esta etapa pode ser realizada utilizando as componentes com rotação (utilizou-se a rotação “varimax”), a fim de maximizar a interpretação dos loadings dos mesmos.

Loadings:

	RC1	RC2	RC3
nrQuartos	0.723	0.194	
nrWC	0.768		
Sala_estar_m2	0.920	0.108	0.202
lote_m2			0.959
nrAndares	0.506	-0.606	
vista	0.296	0.389	0.139
condicao	-0.126	0.591	
piso_m2	0.821	-0.321	0.295
arrecadacao_m2	0.335	0.801	-0.133

Tabela de loadings das variáveis e componentes principais

RC1: Representa a dimensão interna (área total do interior da casa) da casa;

Este componente serve para capturar a funcionalidade da casa, contendo informação sobre os cômodos e a dimensão de áreas internas.

RC2: Representa o físico estrutural e qualitativo da casa (Quanto mais alta a casa, menor a área da arrecadação);

Reflete o aspeto prático da residência, e contém informação sobre o estado da condição da casa, o tamanho da arrecadação e andares.

RC3: Representa a dimensão externa da casa. (lote_m2)

O último componente captura aspetos externos, como o tamanho do terreno total, o qual tem uma relação com o tamanho interior da residência.

Uma análise que foi essencial para a interpretação da primeira e última componente reside no seguinte fundamento. Num caso geral, é habitual que a dimensão do lote seja superior à do piso, pois referem-se à área total e interior da residência, respetivamente. Contudo, nos casos em que isso não acontece verifica-se que é por conta da altura (andares) que a residência apresenta, tal fenómeno pode ser observado no seguinte output (o código encontra-se tanto no script em R como no anexo 5).

```
"O número registos onde a dimensão do lote é inferior à do piso e o número de andares é igual a 1 é: 0"
"O mesmo caso mas com um número de andares superior a 1 é: 92"
```

3.5. Scores dos Componentes

De modo a finalizar a implementação do PCA foram inseridos os três componentes extraídos ao dataset após o tratamento, pelas de cada componente, para pudermos assim avançar para a secção onde será estudada a implementação do clustering. No dataset as componentes ficaram identificadas como dimensão_interna (primeiro componente), altura (segundo componente) e dimensão_externa (terceiro componente).

4. IDENTIFICAÇÃO DA HETEROGENEIDADE NA BASE DE DADOS

Feita a análise detalhada anterior que permitiu a criação das novas dimensões (componentes principais), procedeu-se com a análise de clustering a fim de ser possível organizar e dividir os dados disponíveis, realizar análises exploratórias sobre os mesmos e ainda realizar previsões baseadas em grupos.

Esta análise de clusters é focada na medida de distâncias, qualificação da semelhança/dissemelhança dos objetos, ou seja, tem como objetivo encontrar grupos de objetos de tal forma que os mesmos presentes em cada grupo sejam semelhantes entre si e diferentes dos demais dos outros grupos, por outras palavras, o desejado é conseguir obter pequenas distâncias entre os elementos dentro dos clusters e distâncias maiores entre os clusters obtidos (non-overlapping).

A aplicação de clustering exige o passo tomado anteriormente, análise de PCA, de modo que não exista a multicolinearidade entre as variáveis (altamente correlacionadas).

4.1. Implementação de Modelos

Para realizar a segmentação dos dados, foram utilizados três métodos de clustering distintos: o **Clustering Hierárquico**, o **K-Means** e o **Modelo Probabilístico (GMM)**. Cada um desses métodos foi avaliado utilizando métricas específicas, permitindo uma comparação da qualidade dos clusters gerados.

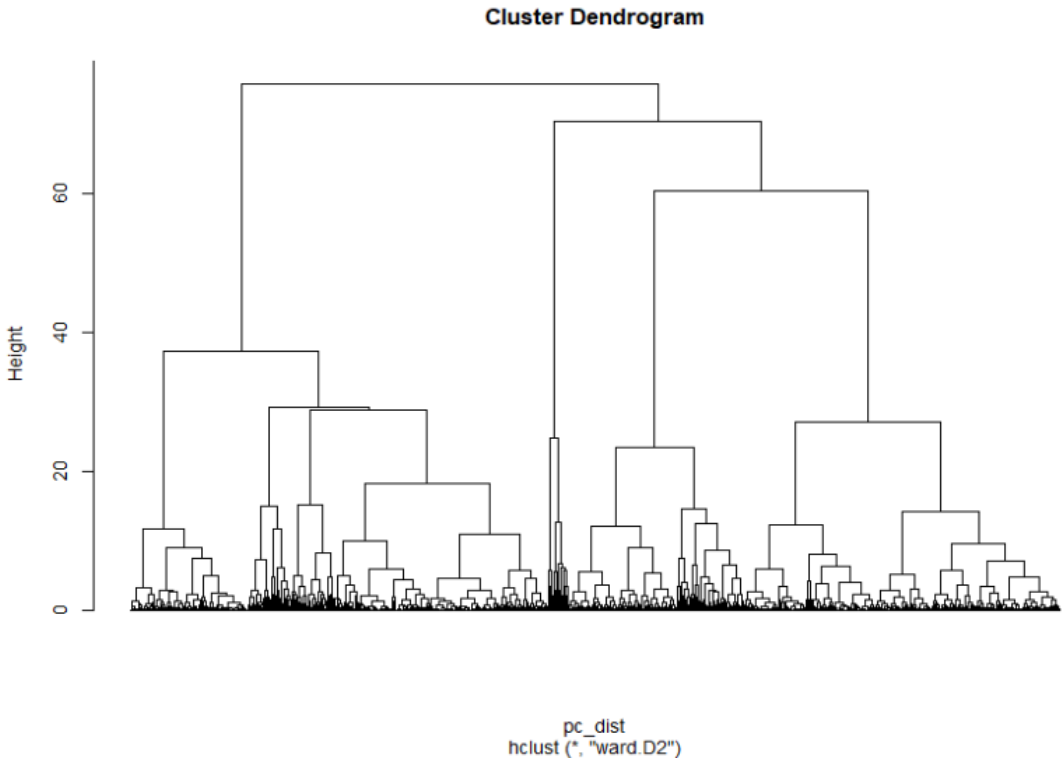
4.1.1. Clustering hierárquico

Nesta análise, utilizou-se o clustering hierárquico, que organiza os dados em uma estrutura de árvore hierarquica (dendrograma). Uma vantagem deste método é a ausência da necessidade de definir, à priori, o número de clusters, no entanto, o mesmo pode ser sensível a ruídos e outliers. O algoritmo hierárquico tradicional utiliza uma matriz de similaridade ou distância para agrupar os dados. A construção da árvore mencionada pode ser feita através da técnica “Clustering Aglomerative”, que segue as etapas iterativas:

1. Cada ponto inicial é tratado como um cluster.
2. Os dois clusters mais próximos são unidos.
3. A matriz de proximidade é atualizada.

4. Os passos acima são repetidos até restar apenas um cluster que engloba 100% dos dados.

Entre os métodos disponíveis para calcular distâncias entre clusters, utilizou-se o método de Ward, reconhecido por minimizar a variabilidade interna dos grupos.



Com base no dendrograma gerado a partir das componentes do PCA, foi feita uma divisão a uma altura igual a 35, o que resultou em 5 clusters. Esta divisão pode ser visualizada por meio do dendrograma encontrado no anexo 6.

4.1.2. K-Means

Foi depois aplicada a técnica de K-Means, geralmente favorecida para avaliar a qualidade dos clusters pela métrica da Silhueta. O método consiste em dividir as observações em um número pré-definido de clusters, que devem ser determinados antes da análise. Dado que se iniciou propositadamente a análise de clustering pelo método hierárquico, o número de clusters para a aplicação desta técnica já não seria um problema.

4.1.3. Clustering probabilístico

Por fim, o clustering probabilistico, implementado através do Gaussian Mixture Model (GMM), oferece uma abordagem para identificar os agrupamentos dos dados, utilizando distribuições gaussianas para modelar a estrutura dos clusters. Diferentemente

do clustering hierárquico e do K-means, que atribuem cada ponto a um único cluster, o GMM baseia-se na probabilidade de cada ponto pertencer a cada cluster, permitindo uma análise mais refinada, especialmente em casos de sobreposição entre os agrupamentos. Nesta análise, o clustering probabilístico foi aplicado como uma etapa complementar para validar e se possível refinar os agrupamentos obtidos nos métodos anteriores.

4.2. Métricas de Avaliação de Clusters

Tendo implementado os modelos mencionados passou-se para a fase de avaliação da qualidade dos clusters de cada um destes.

No caso do clustering hierárquico, utilizou-se o Método da Silhueta para medir a qualidade da segmentação. O valor obtido foi 0.26 (anexo 7), indicando que, embora os clusters não sejam perfeitamente distintos, há uma separação razoável entre eles.

O K-Means foi avaliado com duas métricas principais. A primeira foi o WCSS (Within-Cluster Sum of Squares), que mede a variação dentro de cada cluster. Este gráfico, encontrado no anexo 8, ajudou a visualizar a dispersão das observações em relação aos centróides dos clusters e foi utilizado para determinar o número adequado de clusters. Em seguida, foi aplicada novamente o Método da Silhueta, que obteve o valor 0.34 (anexo 9), indicando uma separação moderada entre os clusters. Este valor é superior ao do clustering hierárquico, mostrando que o K-Means oferece uma segmentação ligeiramente mais precisa, contudo ainda indica uma estrutura fraca dos clusters, o que pode ser válido dado que os dados iniciais podem não ser reais e, até mesmo se forem, por vezes, é complicado encontrar uma estrutura de clusters forte.

Para o modelo probabilístico, foi utilizado o BIC (Bayesian Information Criterion), que indicou que o melhor modelo teria 9 componentes (anexo 10) com um valor de -23637.65, (anexo 11). Com base nesse critério, foi selecionado o modelo VVV (variable volume, shape and orientation), o que oferece maior flexibilidade na modelagem das variâncias e covariâncias entre os diferentes componentes. Este modelo foi capaz de capturar a complexidade dos dados de maneira mais detalhada, sugerindo que seria o mais adequado para a segmentação. Contudo, como ilustrado pelo anexo 12 os clusters apresentam evidências de sobreposição significativa, comum nestes tipos de modelos, indicando que os clusters não estão completamente bem definidos, exigindo uma cautela na interpretação dos agrupamentos.

Assim foi adicionado ao dataset a variável "cluster", resultado do modelo K-means, que contém o número do cluster ao qual o registo em questão pertence.

4.3. Interpretação dos Clusters

Após realizar os diferentes métodos de clustering, passamos à interpretação dos clusters gerados, levando em consideração as variáveis de perfil: Preço da Residência (*Preco*), Ano de Construção (*Ano_construcao*) e Ano de Renovação/Remodelação (*Ano_renovacao*). A seguir, discutimos a caracterização de cada cluster, baseada nos scores do **PC1**, **PC2** e **PC3** e com os valores obtidos através dos gráficos nos anexos para a interpretação dos clusters:

Cluster 1: Casas Clássicas de Médio Padrão

PC1 (0.29 - médio): Este cluster representa casas com características medianas em relação ao padrão geral do mercado, indicando que não são de luxo, mas ainda possuem um bom valor.

PC2 (1.4 - alto): A pontuação elevada sugere que, apesar do valor médio, essas casas podem ser mais modernizadas ou apresentar qualidades que tornam a sua avaliação superior.

PC3 (-0.3 - ligeiramente baixo): A pontuação negativa indica que essas casas não têm grandes terrenos ou espaços amplos, o que limita sua atratividade em comparação com outras com maior área.

Variáveis de Perfil:

Preço: O preço destas casas varia entre 400-500k, o que as coloca em uma faixa acessível de mercado.

Ano de Construção: Essas casas foram construídas entre 1950-1980, sugerindo um estilo mais antigo e uma infraestrutura que pode precisar de alguma manutenção.

Ano de Renovação/Remodelação: A renovação/ remodelação dessas casas ocorre principalmente em torno de 2000, mas de forma esparsa, o que pode significar melhorias pontuais e não uma renovação contínua.

Cluster 2: Casas Modernas de Alto Padrão

PC1 (1.3 - alto): Este cluster de casas apresenta residências com alta funcionalidade, contendo grandes dimensões de áreas internas.

PC2 (-0.5 - baixo): A pontuação negativa sugere que essas casas podem não ter tantas renovações recentes ou características de renovação em comparação com o resto do mercado. Porém, o alto valor de PC1 compensa este fator.

PC3 (0.01 - médio): A pontuação próxima de zero indica que essas casas não possuem grandes áreas de terreno, o que pode ser um fator limitante, mas, devido ao seu valor e localização, ainda são bem valorizadas.

Variáveis de Perfil:

Preço: O preço destas casas está na faixa de 600-700k, refletindo um mercado mais premium.

Ano de Construção: As casas deste cluster foram construídas pós-1980, mostrando uma infraestrutura mais moderna e maior eficiência nos materiais e design.

Ano de Renovação/Remodelação: Renovação e remodelação ocorreram principalmente a partir de 2000, com poucas renovações após esse período. A modernização parece ter ocorrido de forma intensiva, mas não de forma contínua.

Cluster 3: Casas Tradicionais de Baixo Padrão

PC1 (-0.9 - baixo): Este cluster apresenta casas com valor baixo, refletindo um padrão mais simples e menos valorizado no mercado.

PC2 (0.05 - médio): A pontuação ligeiramente positiva sugere que, apesar do valor baixo, essas casas podem ter algumas renovações ou características que as tornam ligeiramente mais atraentes do que outras de padrão inferior.

PC3 (0.25 - médio): A pontuação média em PC3 indica que essas casas podem ter algum terreno, mas não em grande quantidade, o que as coloca em uma faixa média em termos de espaço disponível.

Variáveis de Perfil:

Preço: O preço dessas casas varia entre 200-300k, o que as coloca na faixa mais acessível do mercado.

Ano de Construção: As casas foram construídas entre 1940-1970, com estilos mais antigos e menos adaptadas às exigências modernas.

Ano de Renovação/Remodelação: O aumento gradual nas renovações indica uma tentativa de modernização ao longo do tempo, mas sem grandes transformações.

Cluster 4: Casas Modernas e de Médio Padrão

PC1 (0.05 - médio): Estas casas têm um valor médio em termos de características gerais, sem grandes diferenciais, mas ainda atendem a um público-alvo que busca conforto e funcionalidade.

PC2 (-0.85 - baixo): A pontuação bastante negativa em PC2 reflete que, em termos de modernização, estas casas podem ter sofrido poucas reformas ou apresentam características mais simples.

PC3 (-0.45 - baixo): A pontuação negativa de PC3 sugere que essas casas também não possuem grandes terrenos ou espaços amplos, limitando sua atratividade.

Variáveis de Perfil:

Preço: O preço dessas casas está na faixa de 400-500k, o que as coloca em uma faixa de mercado intermediária.

Ano de Construção: Estas casas foram construídas entre 1980-2020, refletindo uma arquitetura mais moderna, mas talvez não tão inovadora.

Ano de Renovação/Remodelação: Renovação e remodelação ocorreram principalmente em 2000, com um aumento gradual depois disso.

Cluster 5: Casas de Luxo / Residências Rurais

PC1 (1 - alto): A pontuação alta indica que têm um valor muito elevado devido às suas características exclusivas e de alto padrão.

PC2 (-0.15 - médio baixo): A pontuação ligeiramente negativa sugere que essas residências podem não ter sido tão modernizadas quanto as casas de outros clusters.

PC3 (5.21 - extremamente alto): A pontuação extremamente alta reflete os vastos terrenos e grandes propriedades associadas a essas casas, o que as torna extremamente valiosas devido ao espaço e privacidade oferecidos.

Variáveis de Perfil:

Preço: O preço dessas casas pode variar bastante, mas não há um agrupamento claro, o que indica uma grande diferença entre os valores de mercado, possivelmente devido ao tamanho do terreno e qualidade da estrutura.

Ano de Construção: As datas de construção podem variar, mas a ênfase está em características rurais ou de luxo, o que sugere uma mistura de residências antigas e modernas.

Ano de Renovação/Remodelação: A renovação ocorreu principalmente entre 2000, com poucas renovações depois disso. O foco aparenta estar na qualidade da estrutura e no grande terreno.

5. CONCLUSÃO

Este estudo confirma a viabilidade do uso de técnicas de aprendizagem não supervisionada, como o PCA e o clustering hierárquico, para a exploração e segmentação de dados relacionados ao mercado imobiliário. Com o pré-processamento apropriado de variáveis, remoção de outliers e implementação do PCA para reduzir a dimensionalidade sem a perda de significância, os dados usados no estudo conceberam alguns resultados satisfatórios em certas etapas. Os resultados do PCA forneceram três dimensões principais: dimensão interna, estrutura e a dimensão externa. A partir destas, o clustering hierárquico completou a análise para segmentar casas em grupos similares que seguiam padrões semelhantes do mercado imobiliário. No geral, o método utilizado não apenas atendeu à finalidade deste estudo, como também deu margem a várias aplicações futuras, incluindo previsão de preço dos imóveis e descoberta de tendências no mercado.

A análise dos PCA revela que a remodelação e o ano de construção são variáveis fundamentais que impactam diretamente o preço final das casas. As casas construídas após 1980 e com renovação significativa, especialmente em torno de 2000, tendem a ter valores mais altos, refletindo um mercado que prioriza a modernização e a eficiência. Já as casas com anos de construção mais antigos (1940-1970) e pouca renovação mantêm preços mais baixos, com poucas atualizações influenciando sua atratividade no mercado.

Além disso, o tamanho do terreno, particularmente em casas de luxo e rurais, desempenha um papel crucial no valor final da propriedade. A segmentação de casas em grupos de características semelhantes ajuda a entender como as variáveis de renovação e construção moldam o preço final, oferecendo insights valiosos para compradores e investidores no mercado imobiliário.

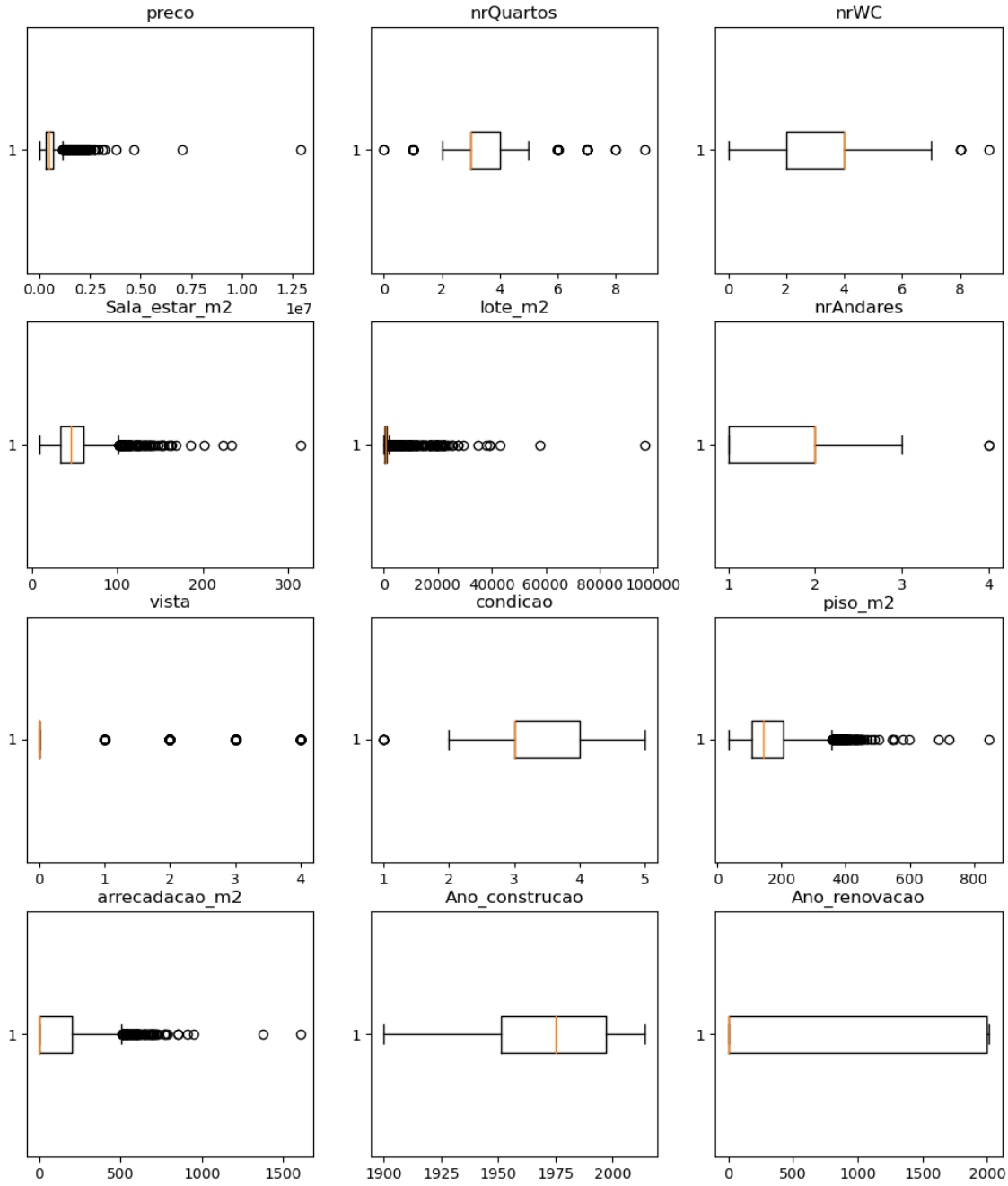
6. BIBLIOGRAFIA

Allen, N. (2024, June 14). What is a water closet and is it the same as a bathroom? Better Homes & Gardens. https://www.bhg.com/what-is-a-water-closet-7499235

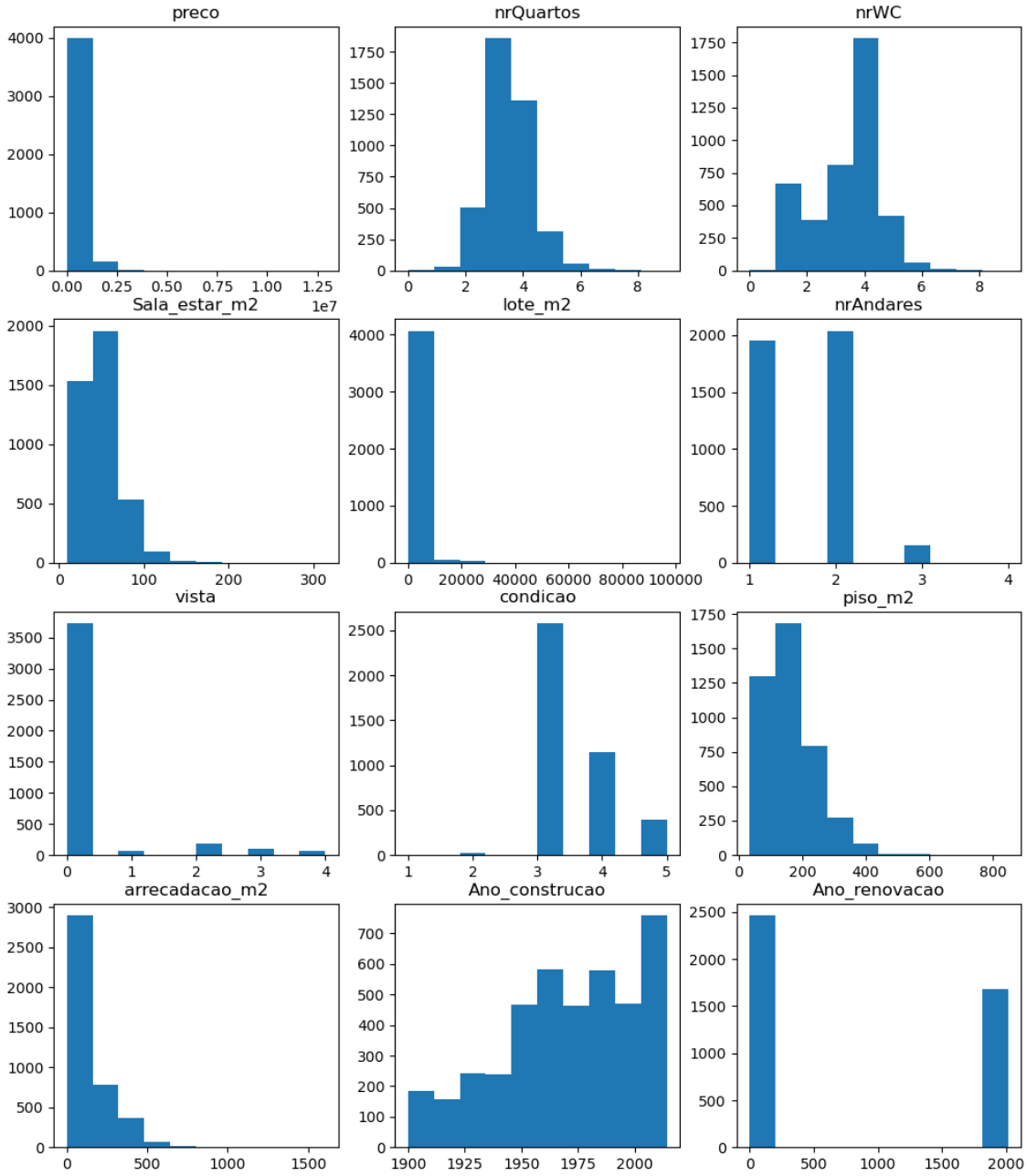
James, Gareth (2013) An introduction to Statistical Learning with Applications in R DOI 10.1007/978-1-4614-7138-7

7. ANEXOS

Anexo 1



Anexo 2



Anexo 3

```
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = correlation)
Overall MSA =  0.53
MSA for each item =
     nrQuartos           nrWC  Sala_estar_m2        lote_m2      nrAndares          vista       condicao       piso_m2
          0.96           0.95           0.48           0.79           0.90           0.92           0.87           0.45
arrecadacao_m2
          0.21
```

Anexo 4

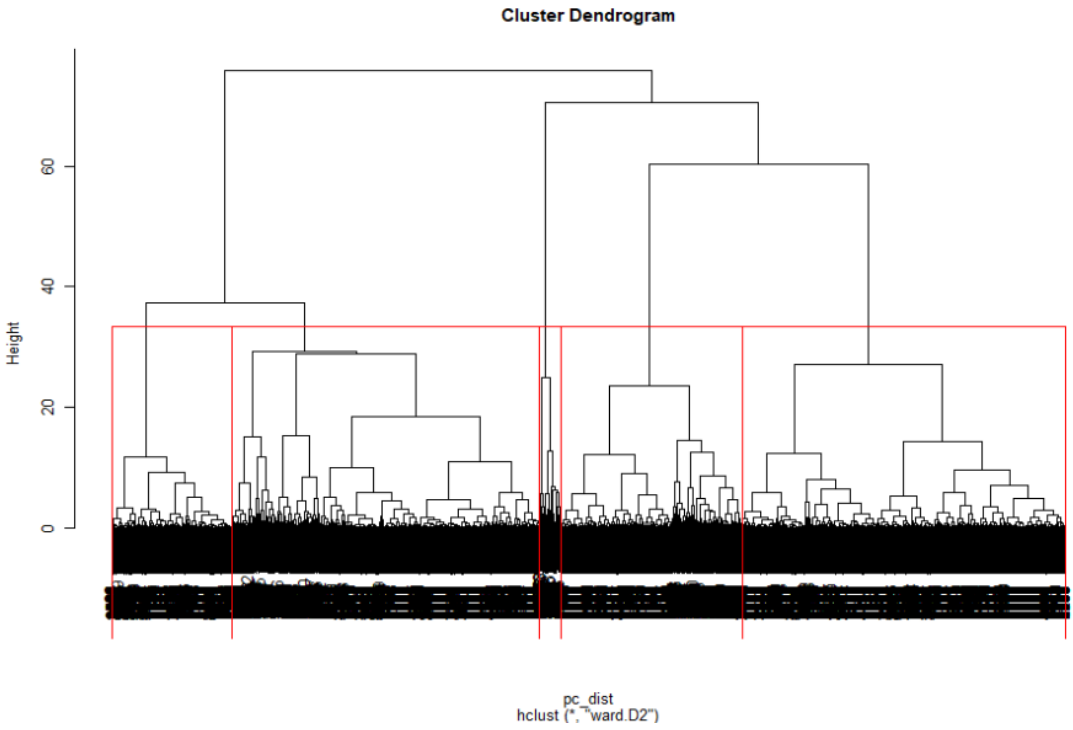
```
     nrQuartos          nrWC  Sala_estar_m2       lote_m2     nrAndares         vista       condicao       piso_m2
     0.5630444     0.5922165     0.8982702     0.9277705     0.6300000     0.2581294     0.3655662     0.8633499
arrecadacao_m2
     0.7716024
```

Anexo 5

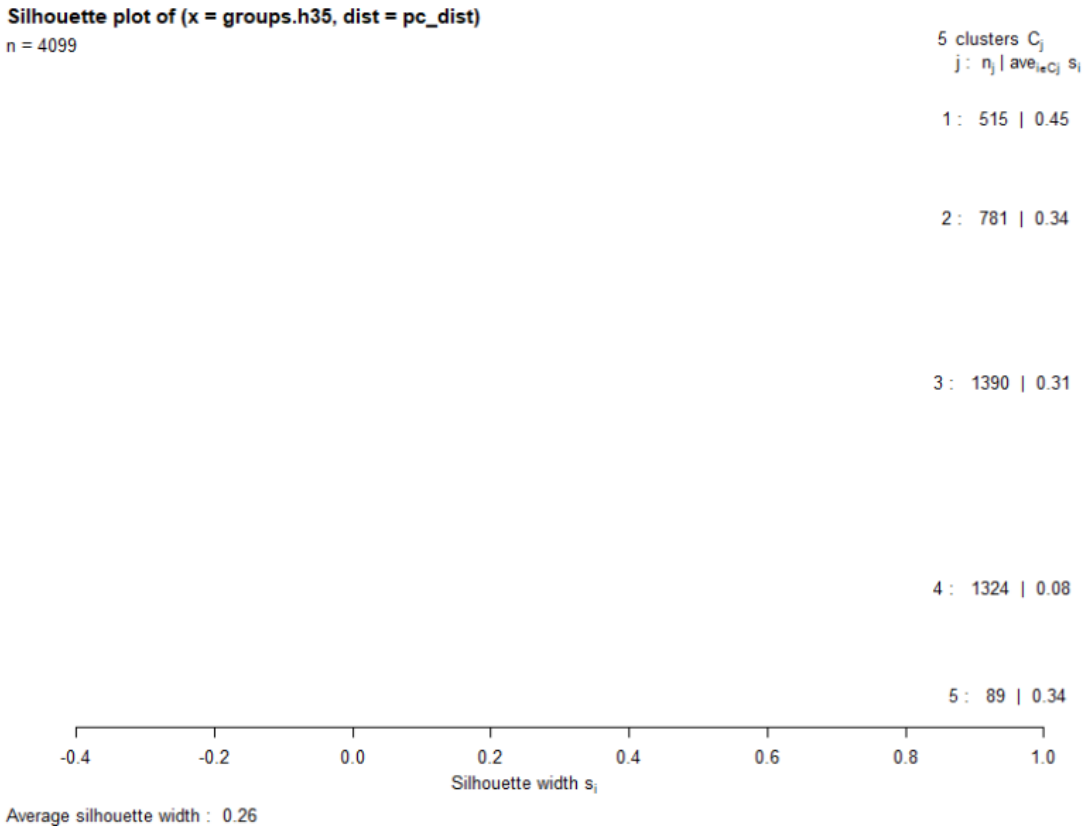
Código sobre análise das variáveis “lote_m2” e “piso_m2”:

paste("O número registos onde a dimensão do lote é inferior à do piso e o número de andares é igual a 1 é:", nrow(dataset[which(dataset$lote_m2 < dataset$piso_m2 & dataset$nrAndares == 1),]))

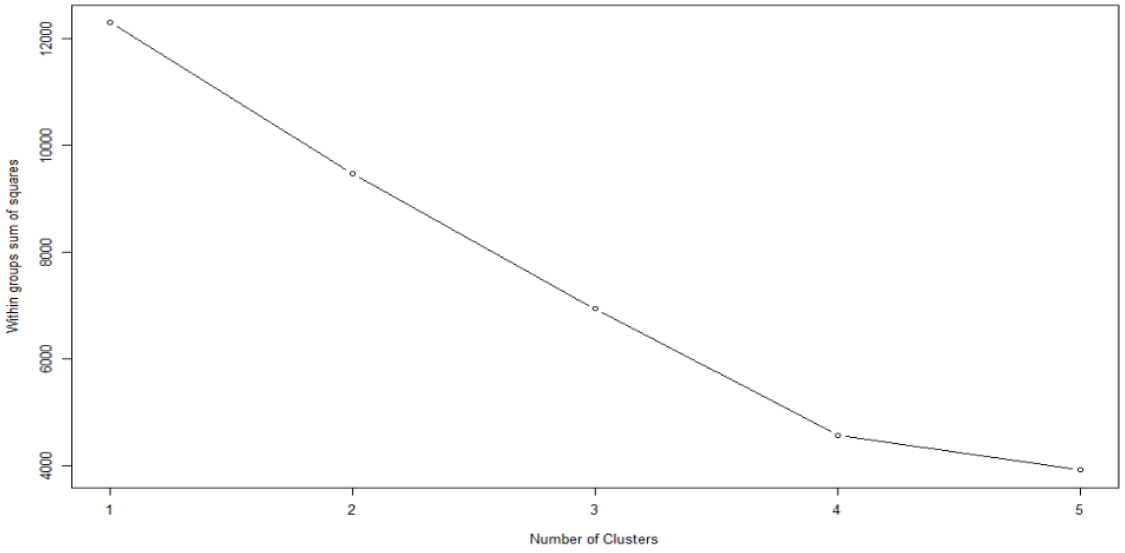
Anexo 6



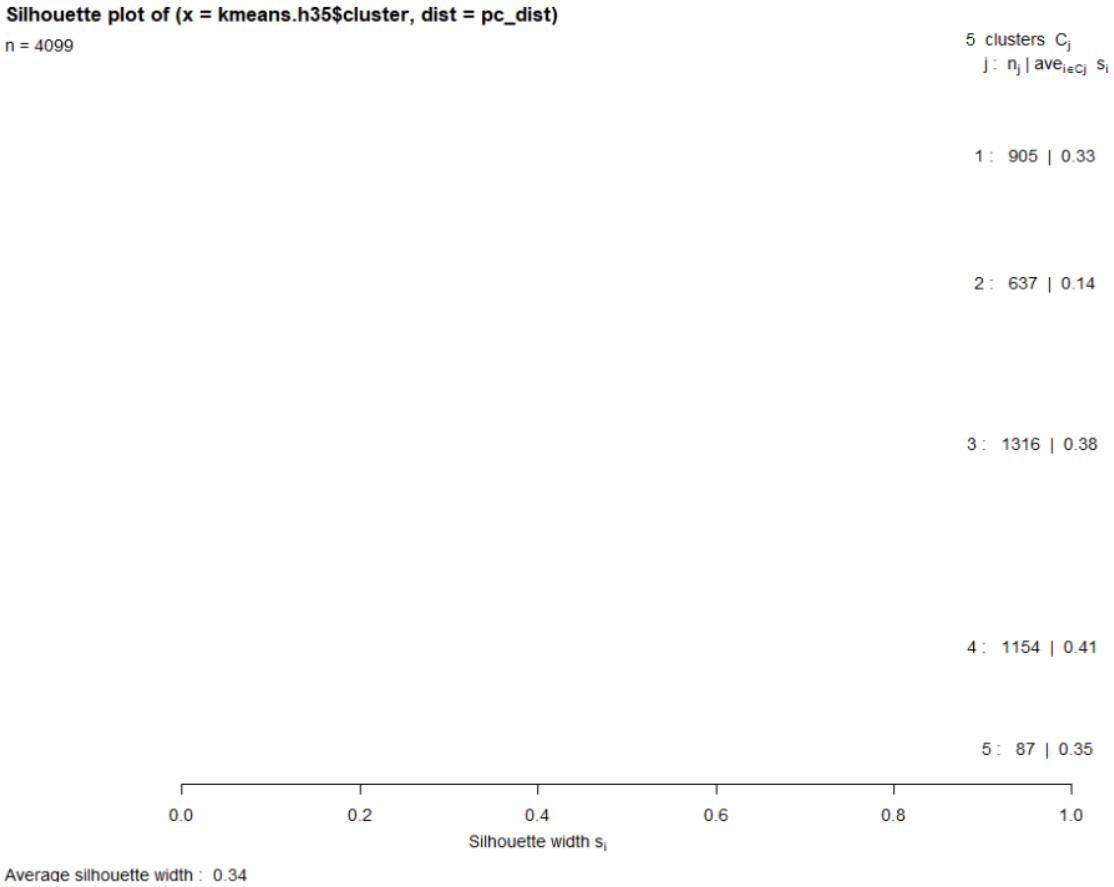
Anexo 7



Anexo 8



Anexo 9



Anexo 10

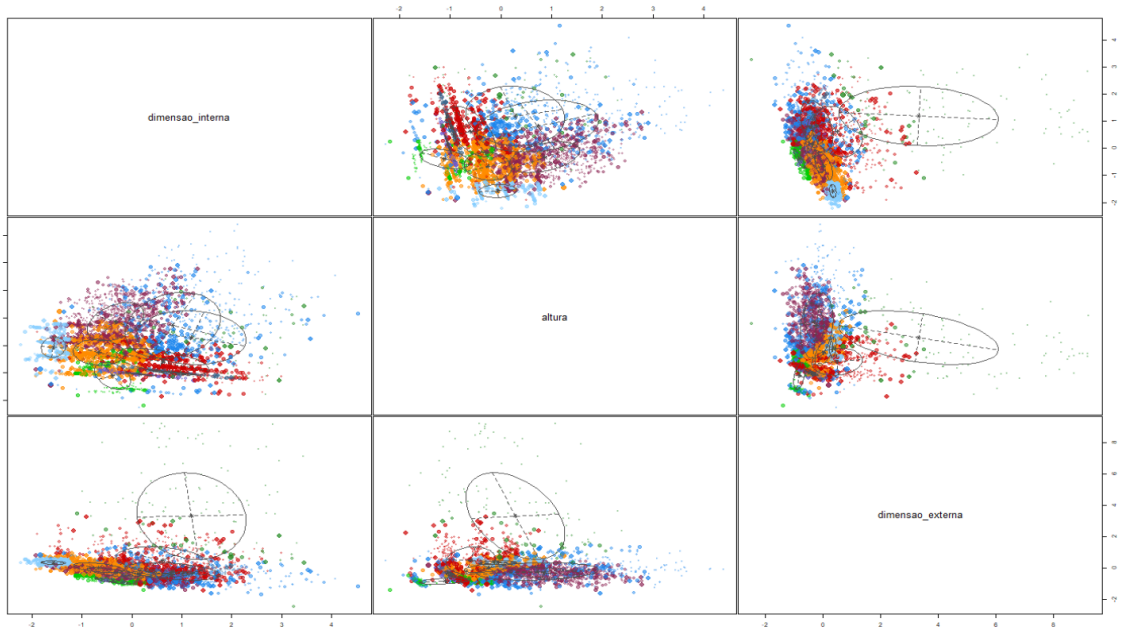
```
-------------------------------------------------------
Gaussian finite mixture model fitted by EM algorithm
-------------------------------------------------------

Mclust VVV (ellipsoidal, varying volume, shape, and orientation) model with 9
components:
```

Anexo 11

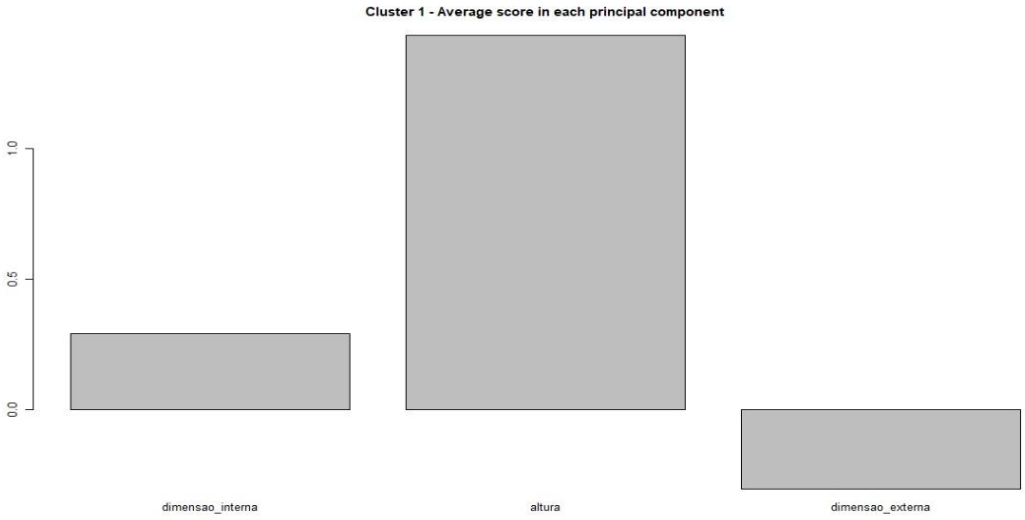
log-likelihood <dbl>	n <int>	df <dbl>	BIC <dbl>	ICL <dbl>
-11448.65	4099	89	-23637.65	-25498.55

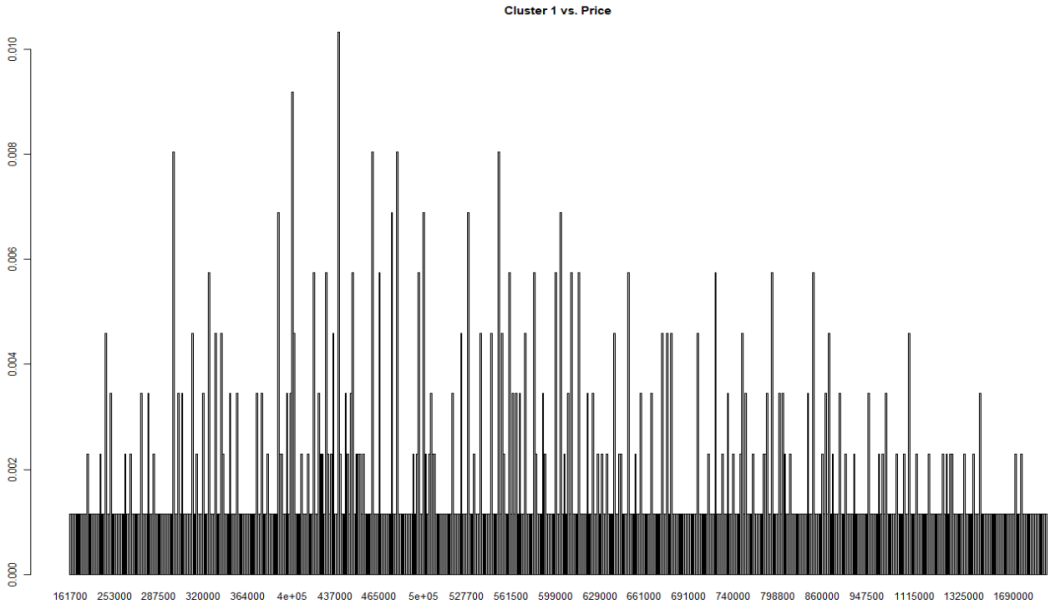
1 row

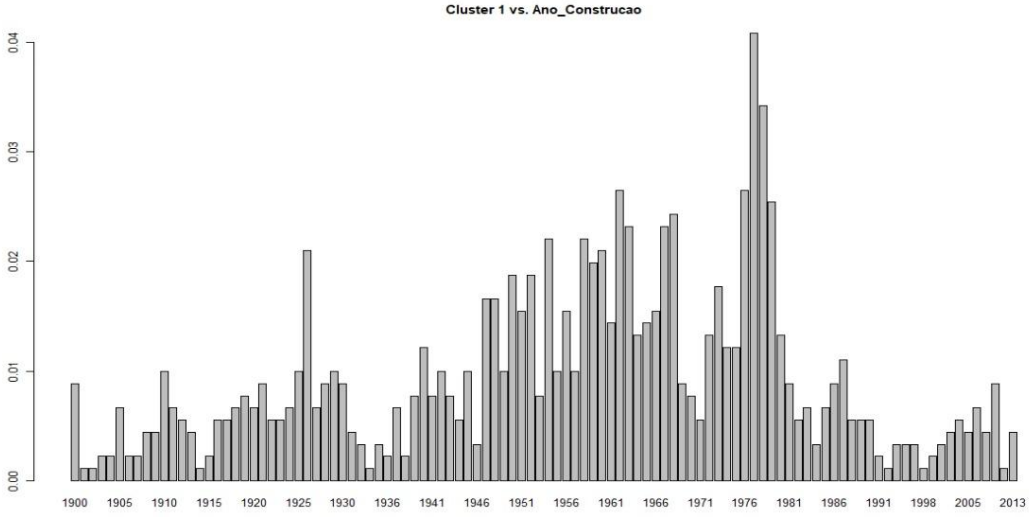
Anexo 12

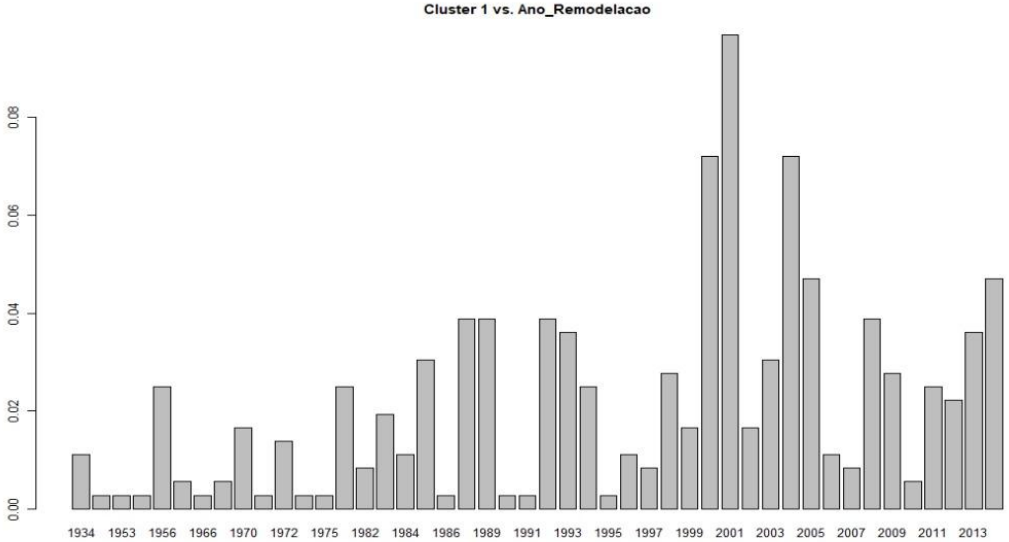


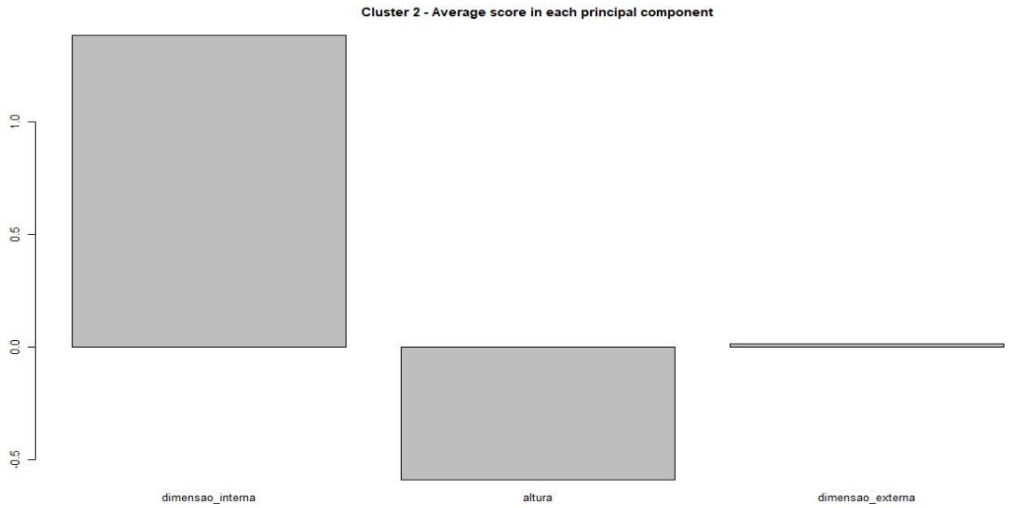
Anexos para a Interpretação dos Clusters

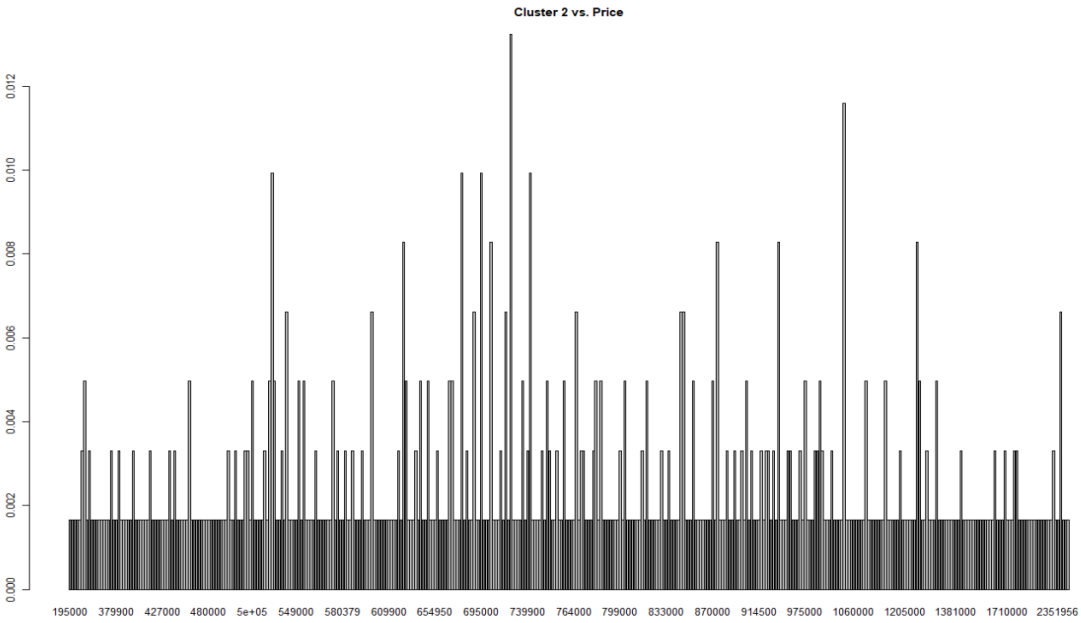


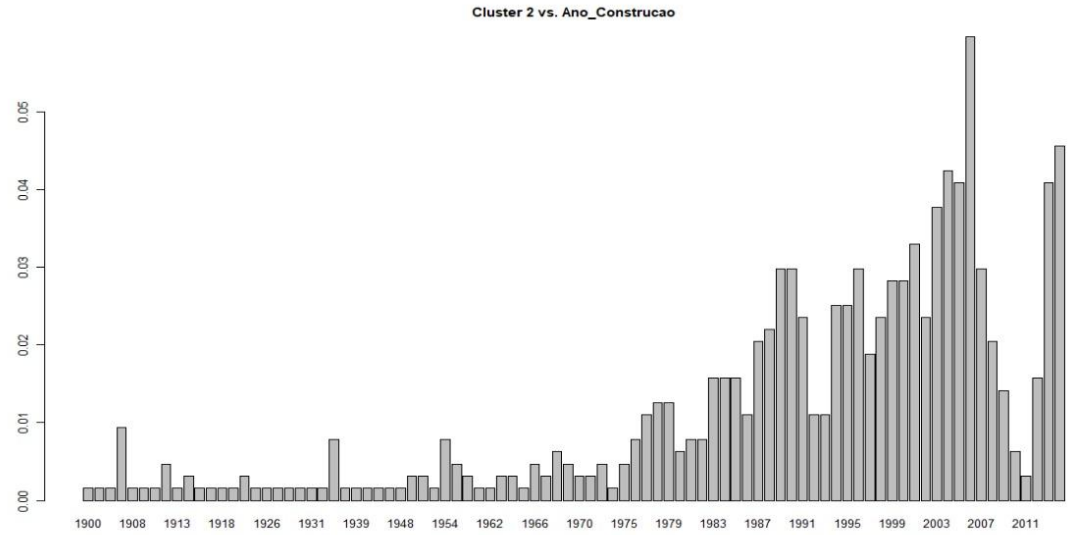

Cluster 1 vs. Price
0.000
0.002
0.004
0.006
0.008
0.010
161700 253000 287500 320000 364000 4e+05 437000 465000 5e+05 527700 561500 599000 629000 661000 691000 740000 798800 860000 947500 1115000 1325000 1690000

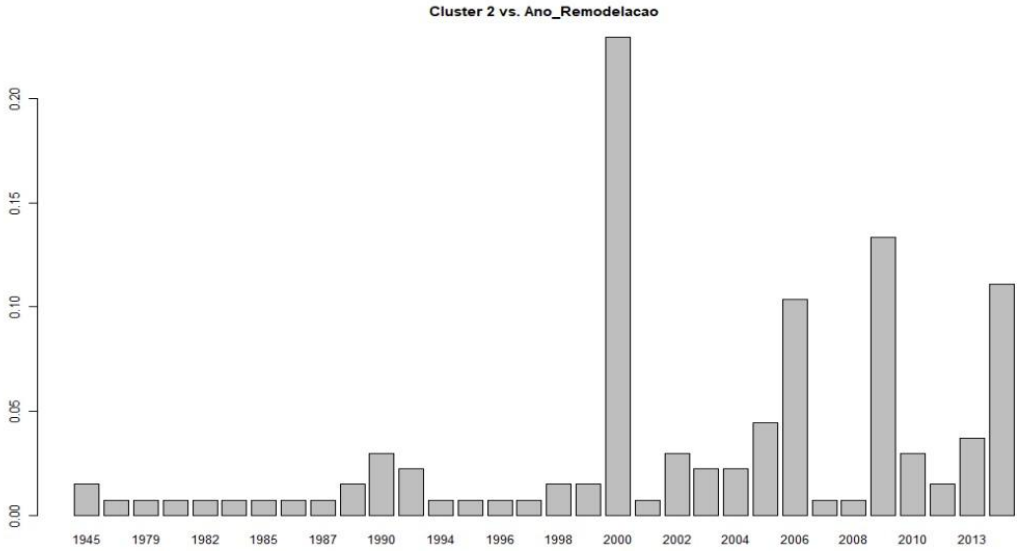

Cluster 1 vs. Ano_Construcao
0.00
0.01
0.02
0.03
0.04
1900 1905 1910 1915 1920 1925 1930 1936 1941 1946 1951 1956 1961 1966 1971 1976 1981 1986 1991 1998 2005 2013

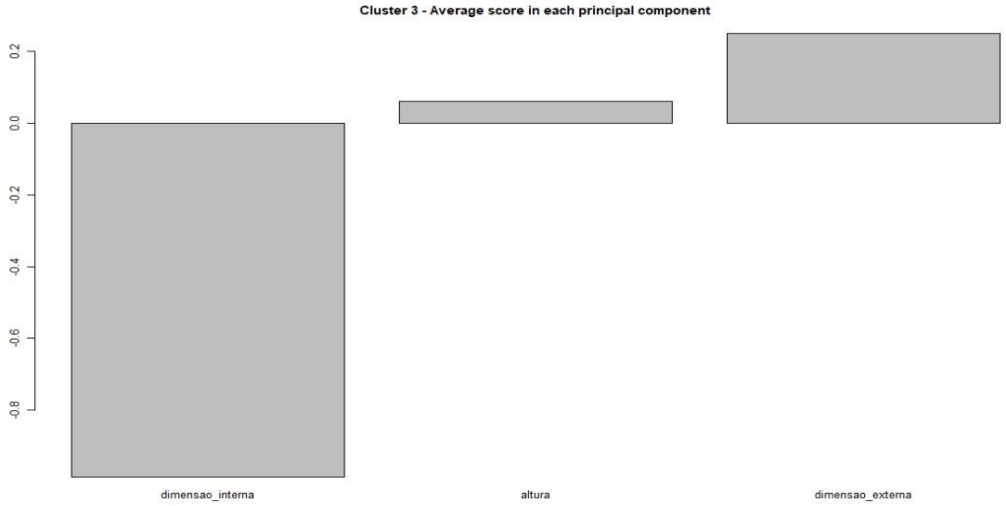

Cluster 1 vs. Ano_Remodelacao
0.00
0.02
0.04
0.06
0.08
1934 1953 1956 1966 1970 1972 1975 1982 1984 1986 1989 1991 1993 1995 1997 1999 2001 2003 2005 2007 2009 2011 2013

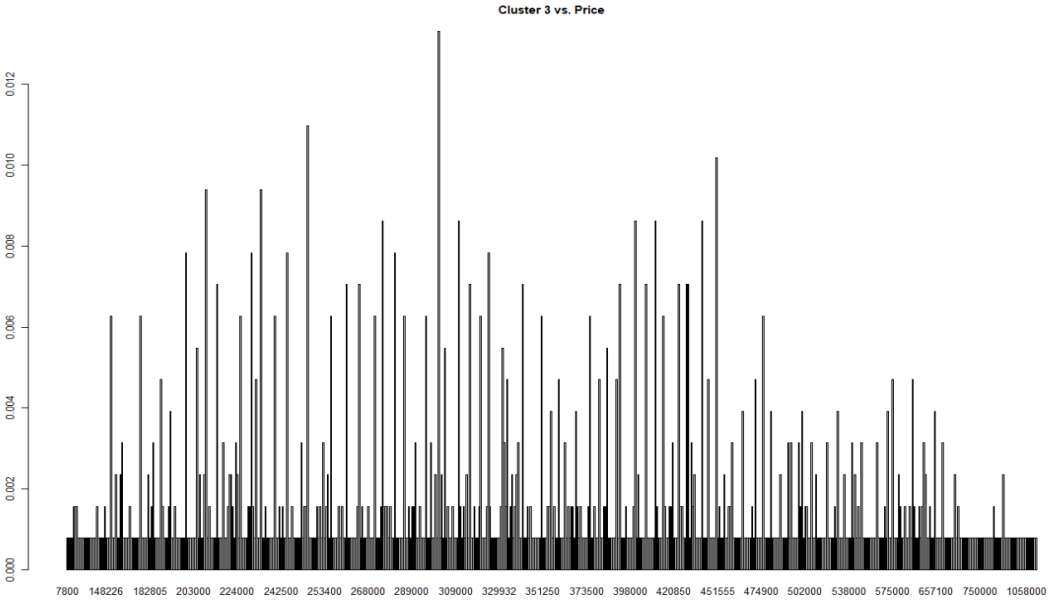

Cluster 2 - Average score in each principal component
1.0
0.5
0.0
-0.5
dimensao_interna
altura
dimensao_externa

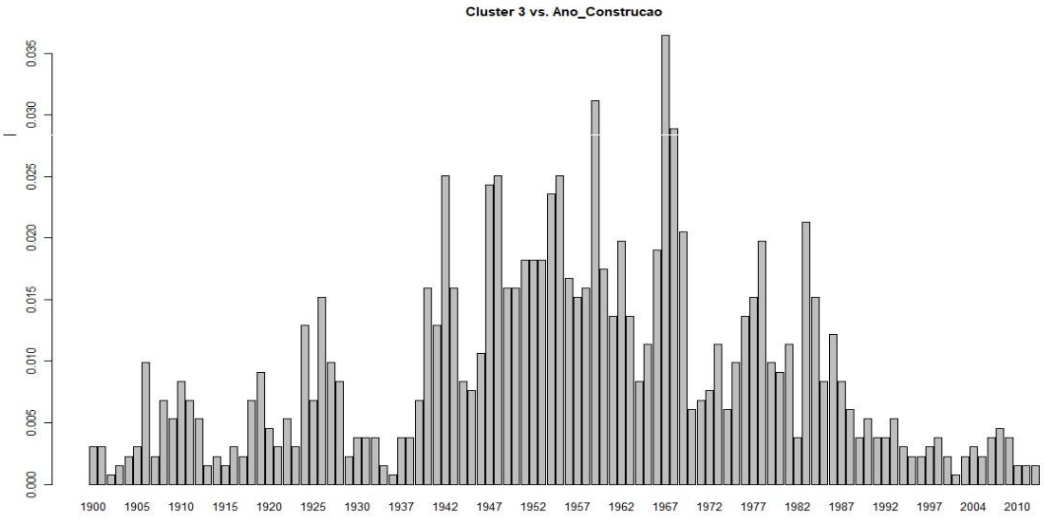

Cluster 2 vs. Price
0.012
0.010
0.008
0.006
0.004
0.002
0.000
195000 379900 427000 480000 5e+05 549000 580379 609900 654950 695000 739900 764000 799000 833000 870000 914500 975000 1060000 1205000 1381000 1710000 2351956

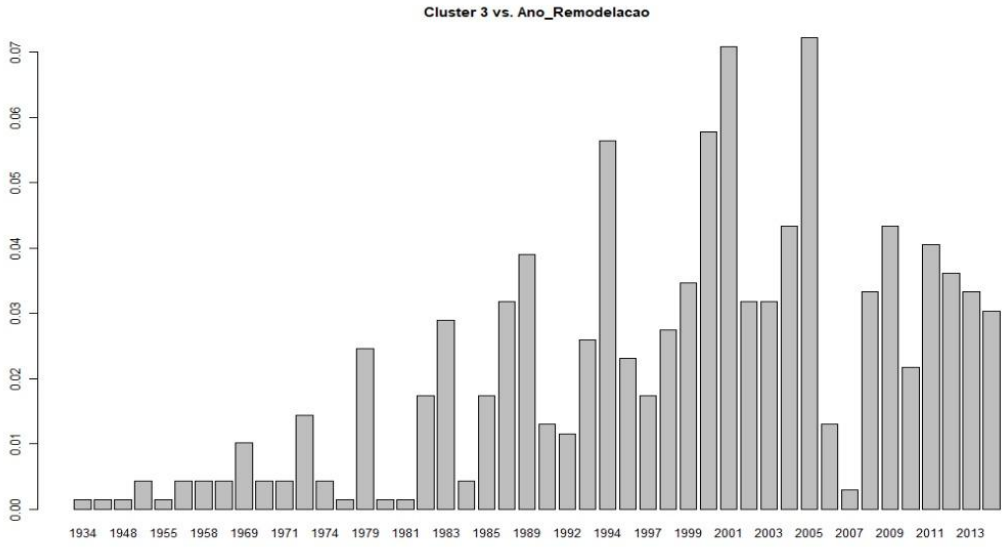

Cluster 2 vs. Ano_Construcao
0.05
0.04
0.03
0.02
0.01
0.00
1900 1908 1913 1918 1926 1931 1939 1948 1954 1962 1966 1970 1975 1979 1983 1987 1991 1995 1999 2003 2007 2011

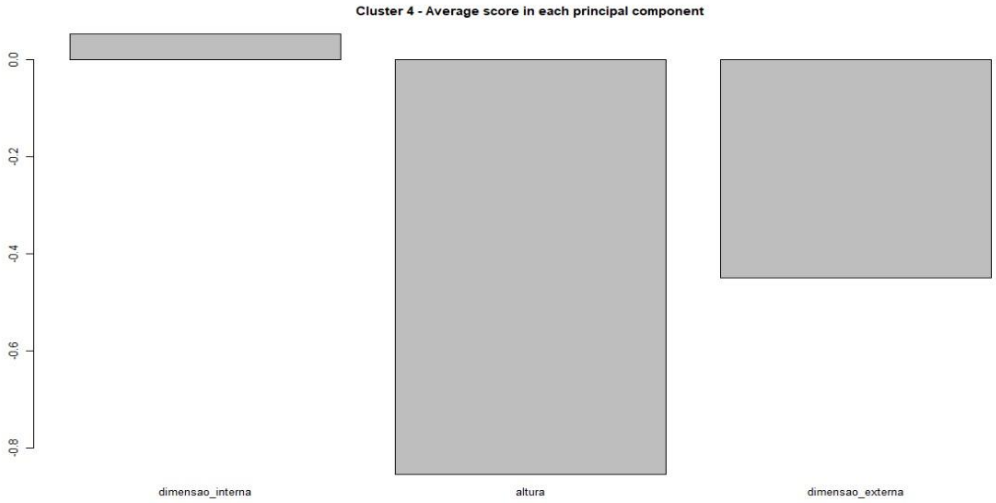

Cluster 2 vs. Ano_Remodelacao
0.00
0.05
0.10
0.15
0.20
1945
1979
1982
1985
1987
1990
1994
1996
1998
2000
2002
2004
2006
2008
2010
2013

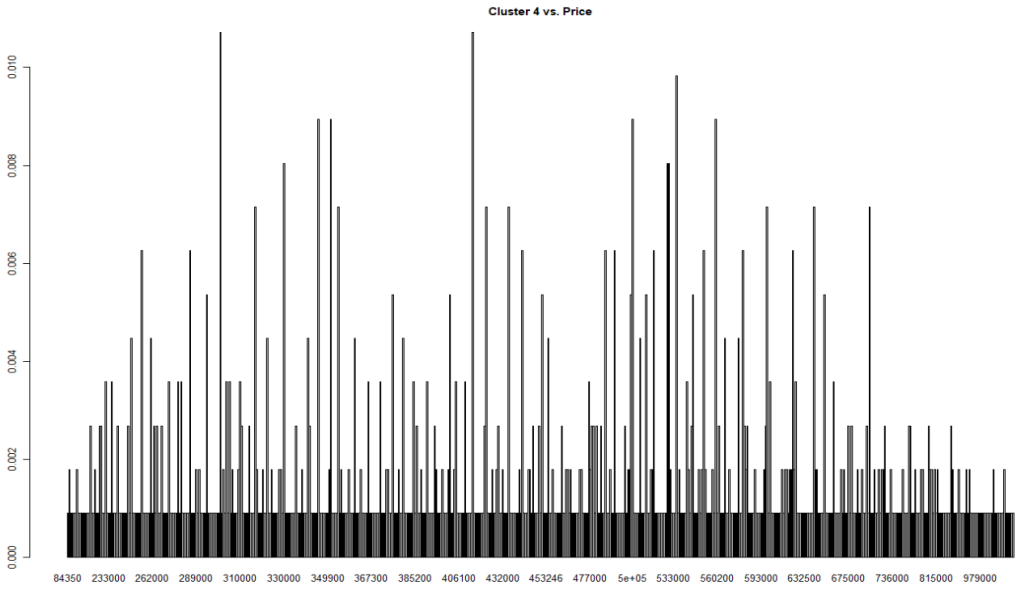

Cluster 3 - Average score in each principal component
-0.8
-0.6
-0.4
-0.2
0.0
0.2
dimensao_interna
altura
dimensao_externa

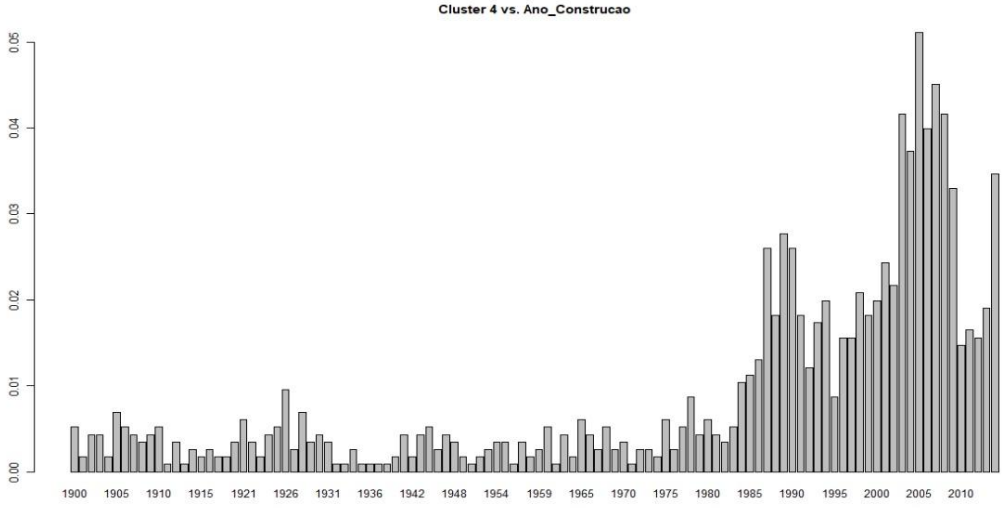

Cluster 3 vs. Price
0.000
0.002
0.004
0.006
0.008
0.010
0.012
7800
148226
182805
203000
224000
242500
253400
268000
289000
309000
329932
351250
373500
398000
420850
451555
474900
502000
538000
575000
657100
750000
1058000

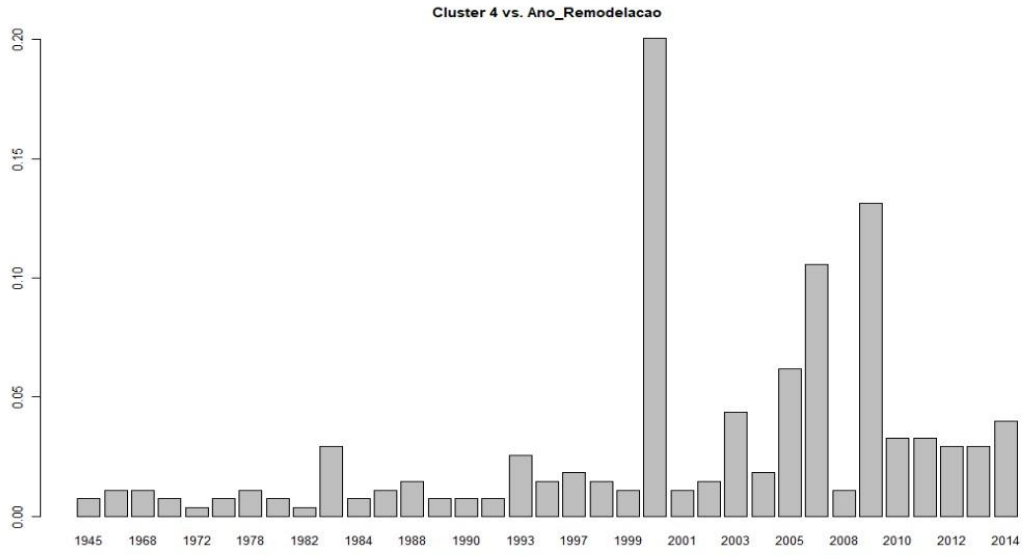

Cluster 3 vs. Ano_Construcao
0.000
0.005
0.010
0.015
0.020
0.025
0.030
0.035
1900 1905 1910 1915 1920 1925 1930 1937 1942 1947 1952 1957 1962 1967 1972 1977 1982 1987 1992 1997 2004 2010

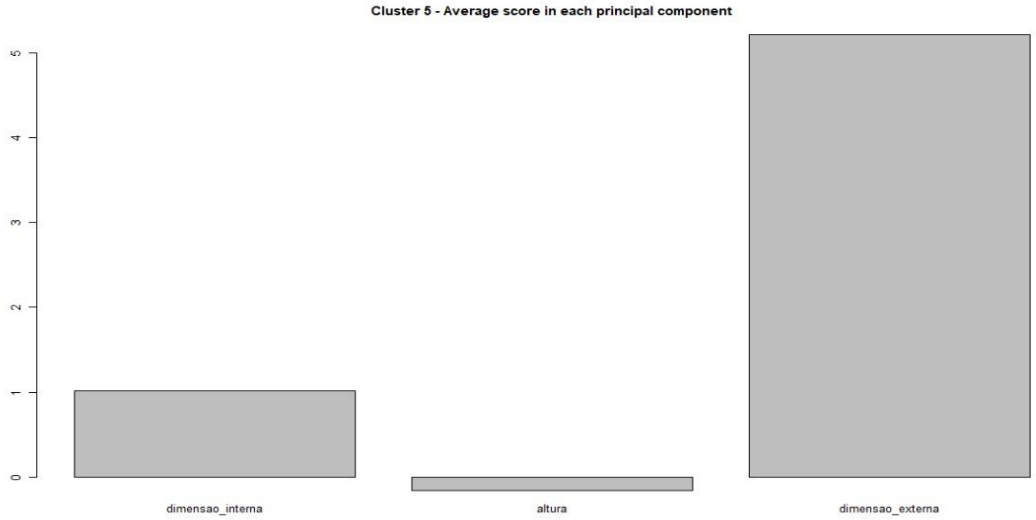

Cluster 3 vs. Ano_Remodelacao
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
1934 1948 1955 1958 1969 1971 1974 1979 1981 1983 1985 1989 1992 1994 1997 1999 2001 2003 2005 2007 2009 2011 2013

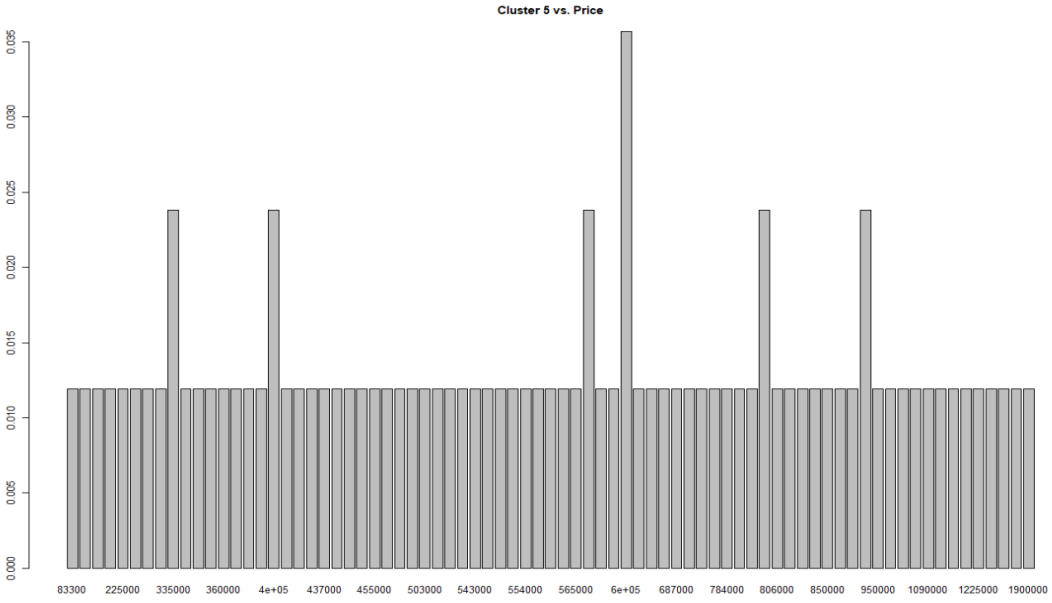

Cluster 4 - Average score in each principal component
0.0
-0.2
-0.4
-0.6
-0.8
dimensao_interna
altura
dimensao_externa

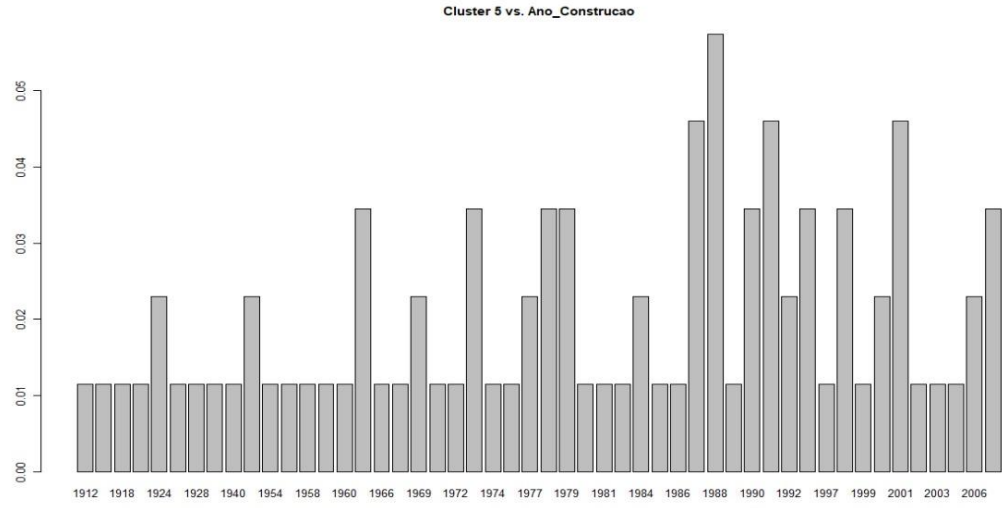

Cluster 4 vs. Price
0.000
0.002
0.004
0.006
0.008
0.010
84350 233000 262000 289000 310000 330000 349900 367300 385200 406100 432000 453246 477000 5e+05 533000 560200 593000 632500 675000 736000 815000 979000

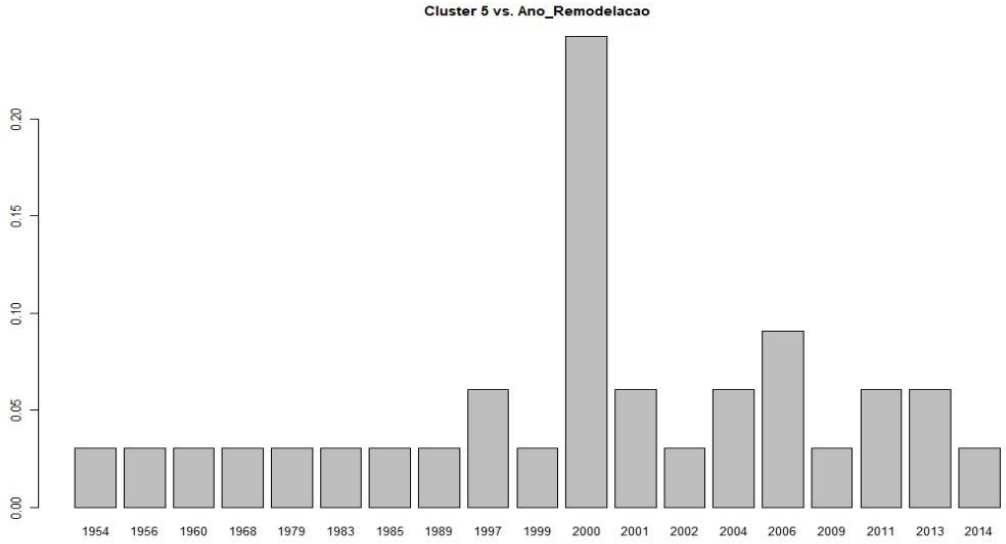

Cluster 4 vs. Ano_Construcao
0.00
0.01
0.02
0.03
0.04
0.05
1900 1905 1910 1915 1921 1926 1931 1936 1942 1948 1954 1959 1965 1970 1975 1980 1985 1990 1995 2000 2005 2010


Cluster 4 vs. Ano_Remodelacao
0.00
0.05
0.10
0.15
0.20
1945 1968 1972 1978 1982 1984 1988 1990 1993 1997 1999 2001 2003 2005 2008 2010 2012 2014


Cluster 5 - Average score in each principal component
0
1
2
3
4
5
dimensao_interna
altura
dimensao_externa


Cluster 5 vs. Price
0.000
0.005
0.010
0.015
0.020
0.025
0.030
0.035
83300
225000
335000
360000
4e+05
437000
455000
503000
543000
554000
565000
6e+05
687000
784000
806000
850000
950000
1090000
1225000
1900000


Cluster 5 vs. Ano_Construcao
0.00
0.01
0.02
0.03
0.04
0.05
1912
1918
1924
1928
1940
1954
1958
1960
1966
1969
1972
1974
1977
1979
1981
1984
1986
1988
1990
1992
1997
1999
2001
2003
2006


Cluster 5 vs. Ano_Remodelacao
0.00
0.05
0.10
0.15
0.20
1954
1956
1960
1968
1979
1983
1985
1989
1997
1999
2000
2001
2002
2004
2006
2009
2011
2013
2014