Alexandre Vasconcelos, N.º 13015
alex-0.5@hotmail.com
Conjunto de Dados - Cinema

RESUMO

O presente projeto consistiu na aplicação de técnicas de análise não-supervisionada a um conjunto de dados, sendo o seu objetivo principal a aquisição de novos e relevantes conhecimentos sobre ele. O trabalho integrou 3 partes, tendo a metodologia seguida abordado as fases de data profiling e de pré-processamento de dados, de clustering e de análise de valores atípicos, possibilitando assim uma exploração aprofundada das características inerentes aos dados e a identificação de padrões e de anomalias significativas.

O conjunto de dados é relativo ao tema *Cinema*, apresentando variáveis como o preço do bilhete, a capacidade da sala de cinema, o horário das sessões, entre outras. A fonte originária deste conjunto aponta como variável-alvo a *Receita_Total*, ou seja, o montante total apurado para uma dada sessão de cinema.

Na Parte 1 do projeto, foi realizado o data profiling de dados, etapa fundamental para a aquisição de conhecimentos acerca do conjunto. Nesse sentido, efetuaram-se diversas análises estatísticas, as quais permitiram revelar distribuições e dependências entre variáveis, e também a necessidade de operar algumas transformações, seguindo técnicas tais como o pré- processamento dos valores omissos e dos valores atípicos, que prepararam os conjuntos de dados para as fases posteriores. Foi igualmente efetuada uma seleção das variáveis redundantes e de baixa variância, tanto para remover informações duplicadas, como para assegurar que apenas as variáveis com capacidade discriminativa seriam consideradas, obtendo-se um conjunto de dados otimizado.

Na Parte 2, teve lugar a tarefa de aprendizagem não-supervisionada de clustering, recorrendo a 2 modelos: K-Means, um modelo de partição, e Aglomerativo, um modelo hierárquico. Inicialmente, foram determinadas as melhores combinações dos parâmetros a utilizar em cada modelo. É de referir que se optou por usar todas as variáveis do conjunto de dados neste processo, pelo que foi necessário realizar uma extração das 2 componentes principais de topo (1 e 2). Os clusters formados e respetivos centroides foram, então, representados num espaço bidimensional das componentes principais. Essa representação foi rigorosamente avaliada mediante tanto índices internos como índices externos. Calcularam-se os centroides, medianas, medoides e dispersão média de cada cluster para fins descritivos. Seguidamente, foi apurada a importância de cada variável para o conjunto de dados. Finalmente, agruparam-se os clusters relativos ao algoritmo K-Means em 5 conjuntos, de acordo com os padrões específicos neles identificados. Foram elaboradas hipóteses para a existência dos desvios detetados e traçados perfis para o conjunto.

Por fim, na Parte 3, foi realizada uma análise de valores atípicos utilizando o modelo de densidade Local Outlier Factor, possibilitando a deteção de anomalias que permitiram adquirir conhecimentos acerca do conjunto de dados. Em primeiro lugar, procedeu-se à identificação das variáveis mais importantes para este processo. De seguida, construíram-se gráficos bivariados tendo em conta todos os pares de variáveis, representando os valores atípicos em questão. Para cada par, foram colocadas algumas hipóteses que poderiam explicar a ocorrência das anomalias. Por fim, recorrendo novamente ao modelo Local Outlier Factor, calcularam-se as pontuações dos valores atípicos e analisou-se a sua distribuição. Para identificar corretamente a quantidade de verdadeiros valores atípicos, utilizou-se o método do Intervalo Interquartil sobre o conjunto das pontuações desses valores.

1. DATA PROFILING E PRÉ-PROCESSAMENTO DE DADOS

1.1. Análise Estatística

Este conjunto é composto por 142.524 registos e 14 variáveis, possuindo um tamanho moderado (Figura d1.1). Todas as variáveis são numéricas, à exceção da variável temporal *Data*, no formato Dia/Mês/Ano (Figura d1.2). Existe uma percentagem ínfima de valores omissos em apenas 2 variáveis (Figura d1.3).

Quanto aos histogramas das variáveis numéricas (Figura d1.4), verificou-se que a maioria deles mostrava uma assimetria à direita, apresentando uma longa cauda que se estendia para valores mais elevados. A maioria dos valores encontrava-se na parte inferior da distribuição. A distribuição log-normal foi a que melhor se ajustou aos dados. Observando o gráfico de barras relativo à componente temporal *Ano*, derivada da variável Temporal *Data* (Figura d1.5), apurou-se que todos os registos se concentravam no ano de 2018.

Utilizando o método do Intervalo Interquartil (com um fator igual a 1.5), verificou-se a existência de uma percentagem significativa de valores atípicos no conjunto de dados (Figura d1.6), nomeadamente nas variáveis *Receita_Total* (11%), *Numero_Bilhetes_Vendidos* (10%), *Numero_Bilhetes_Utilizados* (10%), *Capacidade_Sala* (8%), *Percentagem_Ocupacao_Sala* (6%) e *Hora* (5%). Esta última evidência foi confirmada pela visualização dos diagramas de caixa e bigodes (Figura d1.7), que permitiram ainda comprovar a existência de várias distribuições assimétricas positivas, como, por exemplo, para a variável *Capacidade_Sala*. Verifica-se, então, a presença de valores atípicos correspondentes a valores muitíssimo altos.

O conjunto dos gráficos de dispersão presentes na Figura d1.8 mostram que a maior parte dos dados se distribui de forma consideravelmente vasta no domínio das variáveis, isto é, estão bastante dispersos relativamente à média, indicando que as variáveis apresentam, de forma geral, uma variância alta. Existem vários pares de variáveis com uma correlação positiva entre elas, como, por exemplo, os pares *Numero_Bilhetes_Utilizados* e *Receita_Total*, e *Numero_Bilhetes_Vendidos* e *Numeros_Bilhetes_Utilizados*.

Após a codificação da variável temporal *Data* (a única a codificar) através da técnica de extração de componentes temporais, verificou-se que a nova variável *Ano* apresentava aproximadamente 60% de valores omissos, sendo que esta última variável apresentava o mesmo valor para todos os registos (2018); uma vez que tinha uma variância nula, decidiu-se removê-la do conjunto de dados. Dada a percentagem ínfima de valores omissos no restante conjunto de dados, optou-se por remover todos os registos com valores omissos, passando o conjunto de dados a mostrar 142.399 registos.

Quanto aos valores atípicos, foi efetuado um truncamento aos limites inferior e superior utilizando o método do Intervalo Interquartil (com um fator igual a 1.5), o que possibilitou a remoção da grande maioria desses valores.

1.2. Relevância e Dependência das Variáveis

Foi construído um mapa de calor de correlação entre as variáveis (Figura d1.9), constatando-se que as dos pares *Numero_Bilhetes_Utilizados* vs *Numero_Bilhetes_Vendidos*, *Numero_Bilhetes_Utilizados* vs *Receita_Total*, *Numero_Bilhetes_Vendidos* vs *Receita_Total*, e *Mes* vs *Trimestre*, evidenciaram uma correlação muito forte entre si (superior a 0.90), sendo, por isso, redundantes, pelo que uma variável de cada par foi removida (*Numero_Bilhetes_Utilizados*, *Numero_Bilhetes_Vendidos* e *Trimestre*). Esta remoção teve por base a permanência da variável-alvo *Receita_Total* no conjunto de dados, justificando-se, no último par, pela maior granularidade que a variável *Mes* oferece, permitindo análises temporais mais detalhadas. De seguida, apurou-se a variância das restantes variáveis, verificando-se que a da *Numero_Bilhetes_Cancelados* era nula, sendo removida do conjunto de dados. As restantes variáveis apresentavam uma variância superior ao limite mínimo pré-definido (0.10), pelo que foram mantidas.

Por fim, dado que a variável-alvo *Receita_Total* apresentava valores numéricos, procedeu-se à sua discretização em 3 intervalos de tamanho muito semelhante, cujas classes foram designadas por *Baixa*, *Media* e *Alta* (Figura d1.10). No conjunto de dados pós-discretização, esta variável passou a chamar-se *Receita_Total_Discretizada*.

2. CLUSTERING

A tarefa de clustering teve como principal objetivo encontrar padrões subjacentes aos dados, sem conhecimentos prévios, quer da importância de cada variável, quer das relações entre elas. Para tal, foram utilizados 2 métodos de clustering distintos, K-Means (partição) e Aglomerativo (hierárquico). Deste modo, optou-se por incluir no clustering todas as variáveis do conjunto de dados total, o que permitiu identificar padrões entre elas, omitidos no caso de haver uma seleção prévia. A inclusão de todas assegurou que nenhuma informação relevante fosse descartada, possibilitando uma análise mais completa. Devido às limitações computacionais impostas pelo elevado número de registos, optou-se por utilizar uma amostra de 30% do conjunto de dados total, reduzindo o tempo de computação e melhorando a eficiência, sem comprometer a representatividade da análise.

2.1 Distâncias e Métodos

No algoritmo K-Means, uma vez que a distância padrão utilizada é a euclidiana, foi tida em conta apenas esta medida de distância para a formação dos clusters. Para o algoritmo Hierárquico Aglomerativo, foram testadas múltiplas combinações de 3 parâmetros: número de clusters (intervalo 10-15), critério de ligação (mínima, média, máxima e de ward) e medida de distância (euclidiana, manhattan e cosseno), tendo sido calculado o respetivo índice de silhueta global. Deste modo, foi selecionada a combinação de parâmetros que maximizou o índice de silhueta global, com valor igual a 0.762, ou seja, um número ótimo de clusters igual a 12, critério de ligação de ward e distância euclidiana (Tabelas d1.1 e d1.2).

Tabela d1.1 - Algoritmo de Clustering Hierárquico Aglomerativo - Índice de Silhueta Global em função do Número de Clusters (k) e do Critério de Ligação.

Conjunto de Dados 1 - Algoritmo de Clustering Hierárquico Aglomerativo		
Número de Clusters (k)	**Critério de Ligação**	**Índice de Silhueta Global**
10	Mínima	0.097
	Média	0.657
	Máxima	0.637
	Ward	0.753
11	Mínima	0.097
	Média	0.656
	Máxima	0.654
	Ward	0.755
12	Mínima	0.097
	Média	0.654
	Máxima	0.654
	Ward	0.762
13	Mínima	0.097
	Média	0.712
	Máxima	0.658
	Ward	0.757
14	Mínima	0.097
	Média	0.718
	Máxima	0.611
	Ward	0.756
15	Mínima	-0.113
	Média	0.728
	Máxima	0.617
	Ward	0.752

Tabela d1.2 - Algoritmo de Clustering Hierárquico Aglomerativo - Índice de Silhueta Global em função do Número de Clusters (12), Critério de Ligação de Ward e Medida de Distância (Euclidiana, Manhattan e Cosseno).

Conjunto de Dados 1 - Algoritmo de Clustering Hierárquico Aglomerativo			
Número de Clusters (k)	Critério de Ligação	Distância	Índice de Silhueta Global
12	Ward	Euclidiana	0.762
		Manhattan	0.740
		Cosseno	-0.122

2.2 Número de Clusters

Para o algoritmo K-Means, de forma a selecionar o número ótimo de clusters, optou-se por calcular o índice de silhueta global para o intervalo de número de clusters 2 -15 (Tabela d1.3). Por consequência, foi selecionado um número ótimo de clusters igual a 14, para maximizar o índice de silhueta global (0.773).

Tabela d1.3 - Algoritmo de Clustering K-Means - Índice de Silhueta Global em função do Número de Clusters (k).

Conjunto de Dados 1 - Algoritmo de Clustering K-Means	
Número de Clusters (k)	Índice de Silhueta Global
2	0.626
3	0.598
4	0.629
5	0.666
6	0.690
7	0.690
8	0.763
9	0.757
10	0.746
11	0.765
12	0.762
13	0.772
14	0.773
15	0.742

Para este algoritmo, foi igualmente calculado o índice interno inércia (soma dos quadrados das distâncias dos pontos aos seus respetivos centroides), obtendo-se um valor aproximado de 134.065.162.036.

Para o algoritmo Hierárquico Aglomerativo foi ainda calculado um outro índice interno, o índice de Davies -Bouldin, tendo sido obtido o valor de 0.135.

2.3 Soluções de Clustering de Referência

Seguidamente, foram calculados os índices externos de pureza e de rand. Para ambos os algoritmos de clustering, K-Means e Hierárquico Aglomerativo, a pureza apresentou um valor de 0.45 e o índice de rand um valor de 0.02.

2.4 Número de Clusters

Para o algoritmo K-Means, de forma a selecionar o número ótimo de clusters, optou-se por calcular o índice de silhueta global para o intervalo de número de clusters 2 -15 (Tabela d1.3). Por consequência, foi selecionado um número ótimo de clusters igual a 14, para maximizar o índice de silhueta global (0.773).

Tabela d1.3 - Algoritmo de Clustering K-Means - Índice de Silhueta Global em função do Número de Clusters (k).

Conjunto de Dados 1 - Algoritmo de Clustering K-Means	
Número de Clusters (k)	Índice de Silhueta Global
2	0.626
3	0.598
4	0.629
5	0.666
6	0.690
7	0.690
8	0.763
9	0.757
10	0.746
11	0.765
12	0.762
13	0.772
14	0.773
15	0.742

Para este algoritmo, foi igualmente calculado o índice interno inércia (soma dos quadrados das distâncias dos pontos aos seus respetivos centroides), obtendo-se um valor aproximado de 134.065.162.036.

Para o algoritmo Hierárquico Aglomerativo foi ainda calculado um outro índice interno, o índice de Davies -Bouldin, tendo sido obtido o valor de 0.135.

2.5 Soluções de Clustering de Referência

Seguidamente, foram calculados os índices externos de pureza e de rand. Para ambos os algoritmos de clustering, K-Means e Hierárquico Aglomerativo, a pureza apresentou um valor de 0.45 e o índice de rand um valor de 0.02.

2.6 Visualização e Descrição

Apenas para efeitos de visualização, optou-se por desenhar o dendrograma referente ao algoritmo Hierárquico Aglomerativo (Figura d1.11), utilizando a melhor combinação de parâmetros previamente definida. Inicialmente, o dendrograma mostra que cada registo é um cluster próprio, num processo ascendente, originando numerosos clusters de menor dimensão na zona inferior. Contudo, ao traçar uma linha de corte a um valor específico da distância euclidiana entre os clusters, os ramos que se unem abaixo dessa linha fundem-se em 12 clusters.

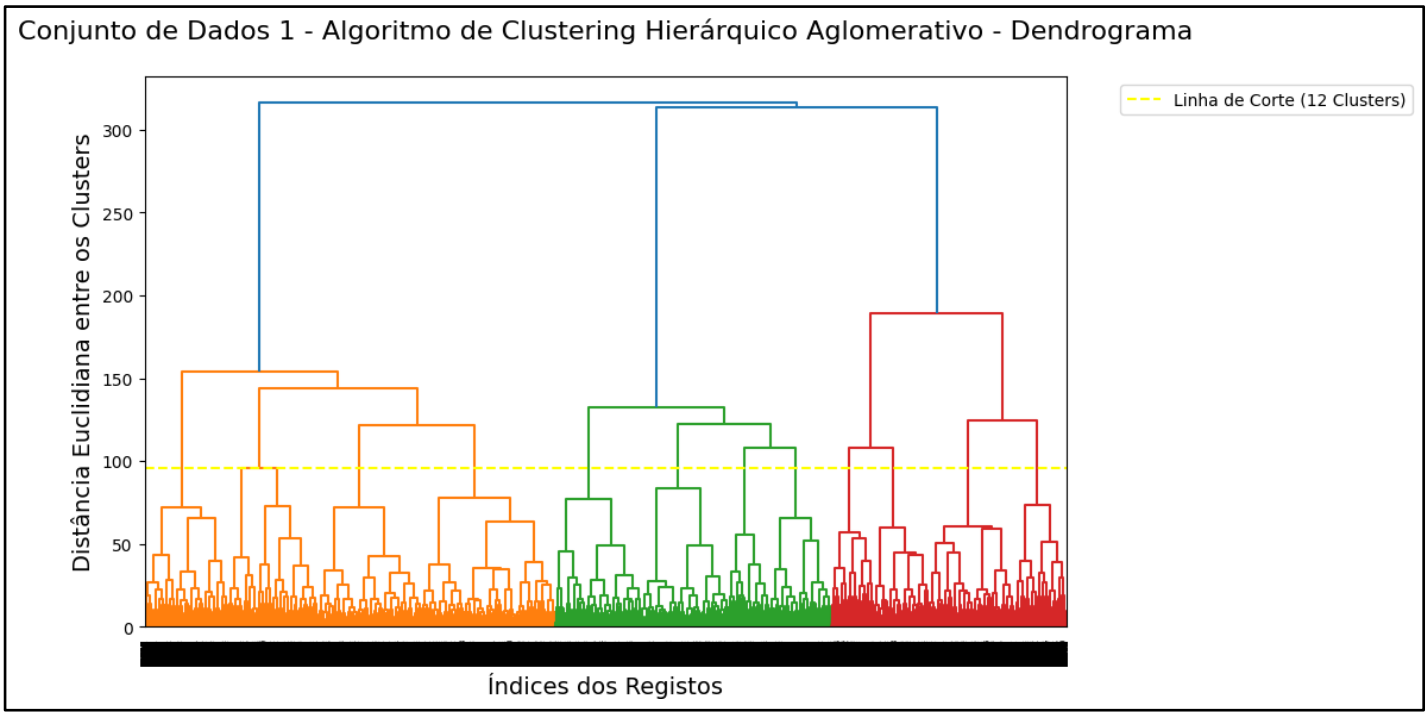

Figura d1.11 - Algoritmo de Clustering Hierárquico Aglomerativo - Dendrograma.

Relativamente à visualização das soluções de clustering mais promissoras, e uma vez que se decidiu não excluir qualquer variável para a tarefa de clustering, optou-se pela extração das componentes principais de topo (1 e 2), de forma a projetar os dados numa representação bidimensional. Na Figura d1.12 é apresentada essa visualização para o K-Means, através da representação dos 14 clusters formados no espaço bidimensional, com os respetivos centróides. Repetiu-se o processo para o Aglomerativo, com 12 clusters representados (Figura d1.13). É de notar que foi efetuada uma normalização de escala apenas para a visualização das soluções de clustering.

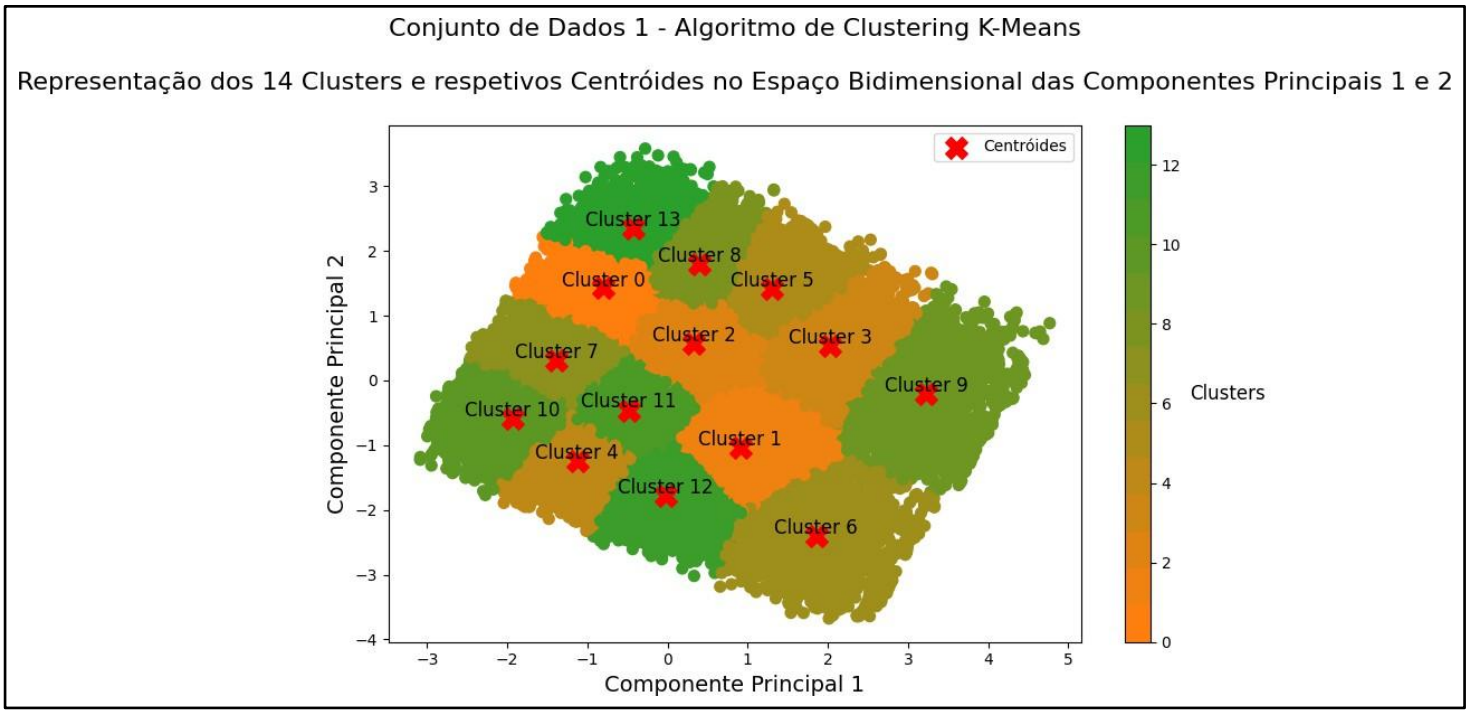

Figura d1.12 - Algoritmo de Clustering K-Means - Representação dos 14 Clusters e respetivos Centróides no Espaço Bidimensional das Componentes Principais 1 e 2.

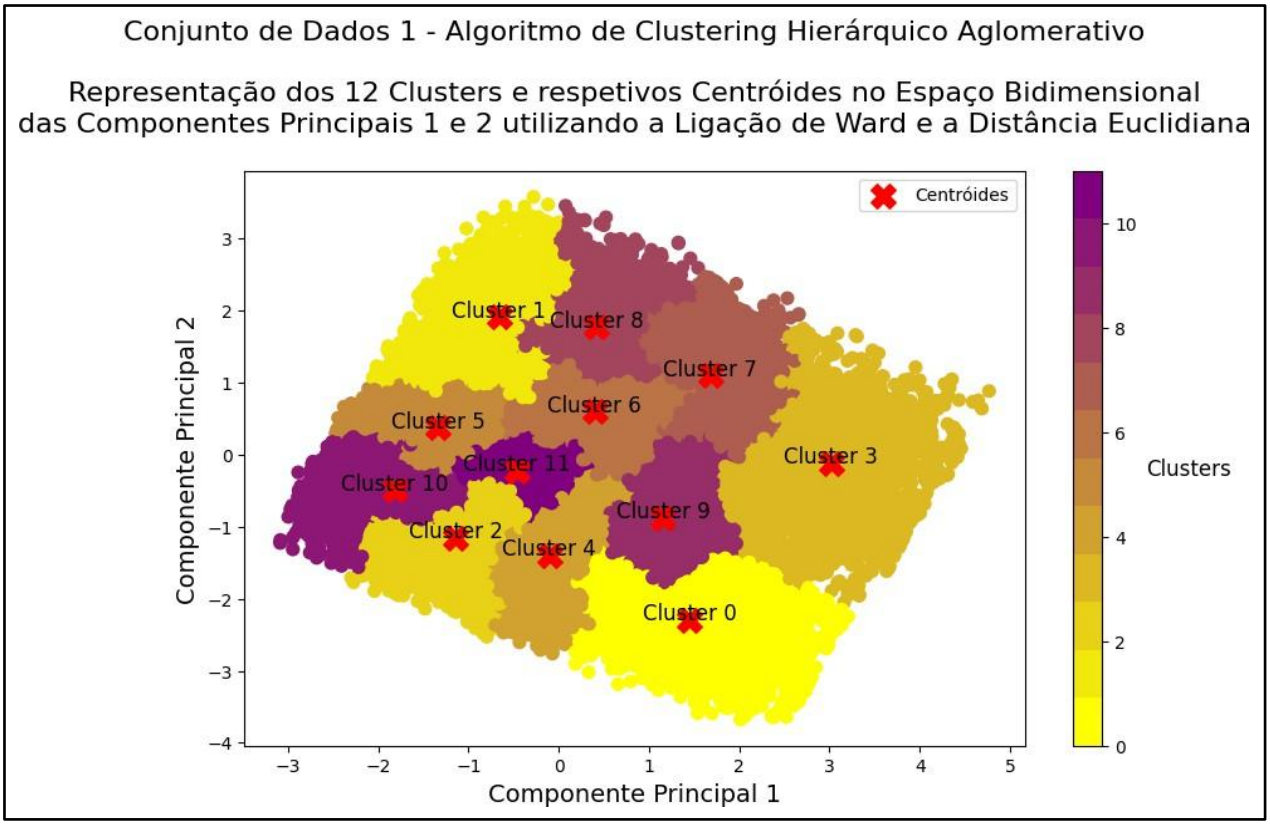

Figura d1.13 - Algoritmo de Clustering Hierárquico Aglomerativo - Representação dos 12 Clusters e respetivos Centróides no Espaço Bidimensional das Componentes Principais 1 e 2 utilizando a Ligação de Ward e a Distância Euclidiana.

Para fins descritivos, relativamente a cada algoritmo, foram calculados os centróides, medianas e medoides para cada cluster, bem como as diferenças entre si; e ainda a dispersão média de cada cluster (Tabelas d1.4 e d1.5).

Tabela d1.4 - Algoritmo de Clustering K-Means - Centróides, Medianas, Medoides e Dispersão Média dos 14 Clusters.

Conjunto de Dados 1 - Algoritmo de Clustering K-Means							
Cluster	Centróide	Mediana	Medoide	Diferença Centróide - Mediana	Diferença Centróide - Medoide	Diferença Mediana - Medoide	Dispersão Média
0	-0.24	-0.27	-0.27	0.03	0.03	0.00	0.32
1	1.10	1.09	1.08	0.01	0.01	0.00	0.49
2	-0.18	0.10	0.06	-0.28	-0.24	0.03	2.20
3	-1.31	-1.47	-1.55	0.16	0.24	0.08	0.72
4	1.56	1.53	1.53	0.03	0.04	0.01	0.50
5	0.34	0.32	0.39	0.02	-0.05	-0.07	1.05
6	-0.70	-0.85	-0.66	0.16	-0.04	-0.19	1.40
7	0.76	0.69	0.68	0.07	0.08	0.01	0.52
8	-0.99	-0.98	-0.93	-0.01	-0.06	-0.05	0.59
9	-1.11	-1.10	-1.10	-0.01	-0.01	-0.00	0.28
10	1.00	1.27	1.19	-0.27	-0.19	0.08	1.31
11	-0.16	-0.13	-0.12	-0.04	-0.05	-0.01	0.81
12	-0.57	-0.54	-0.66	-0.03	0.09	0.12	0.87
13	1.49	1.65	1.52	-0.16	-0.03	0.12	1.65

Tabela d1.5 - Algoritmo de Clustering Hierárquico Aglomerativo - Centróides, Medianas, Medoides e Dispersão Média dos 12 Clusters.

Conjunto de Dados 1 - Algoritmo de Clustering Hierárquico Aglomerativo							
Cluster	Centróide	Mediana	Medoide	Diferença Centróide - Mediana	Diferença Centróide - Medoide	Diferença Mediana - Medoide	Dispersão Média
0	-0.42	-0.57	-0.53	0.15	0.10	-0.05	1.87
1	0.63	0.46	0.23	0.17	0.40	0.23	1.29
2	-1.15	-1.08	-1.08	-0.06	-0.07	-0.01	0.32
3	1.45	1.76	2.00	-0.31	-0.55	-0.25	1.57
4	-0.75	-0.60	-0.51	-0.15	-0.24	-0.09	0.66
5	-0.48	-0.40	-0.21	-0.08	-0.27	-0.20	0.86
6	0.50	0.54	0.54	-0.04	-0.04	-0.00	0.37
7	1.39	1.43	1.44	-0.03	-0.04	-0.01	0.44
8	1.10	1.10	1.11	-0.01	-0.01	-0.01	0.68
9	0.14	0.37	0.59	-0.23	-0.45	-0.22	1.02
10	-1.15	-1.09	-1.06	-0.06	-0.09	-0.03	0.67
11	-0.34	-0.31	-0.31	-0.03	-0.03	-0.00	0.26

A seguir, para os dois algoritmos, foram construídos gráficos de barras (Figuras d1.14 e d1.15) relativos à importância de cada uma das variáveis para a diferenciação de cada cluster formado, através do cálculo do valor F do teste de Análise de Variância (ANOVA). Estes gráficos incluem também uma representação final referente à importância global de todas as variáveis para a separação dos clusters.

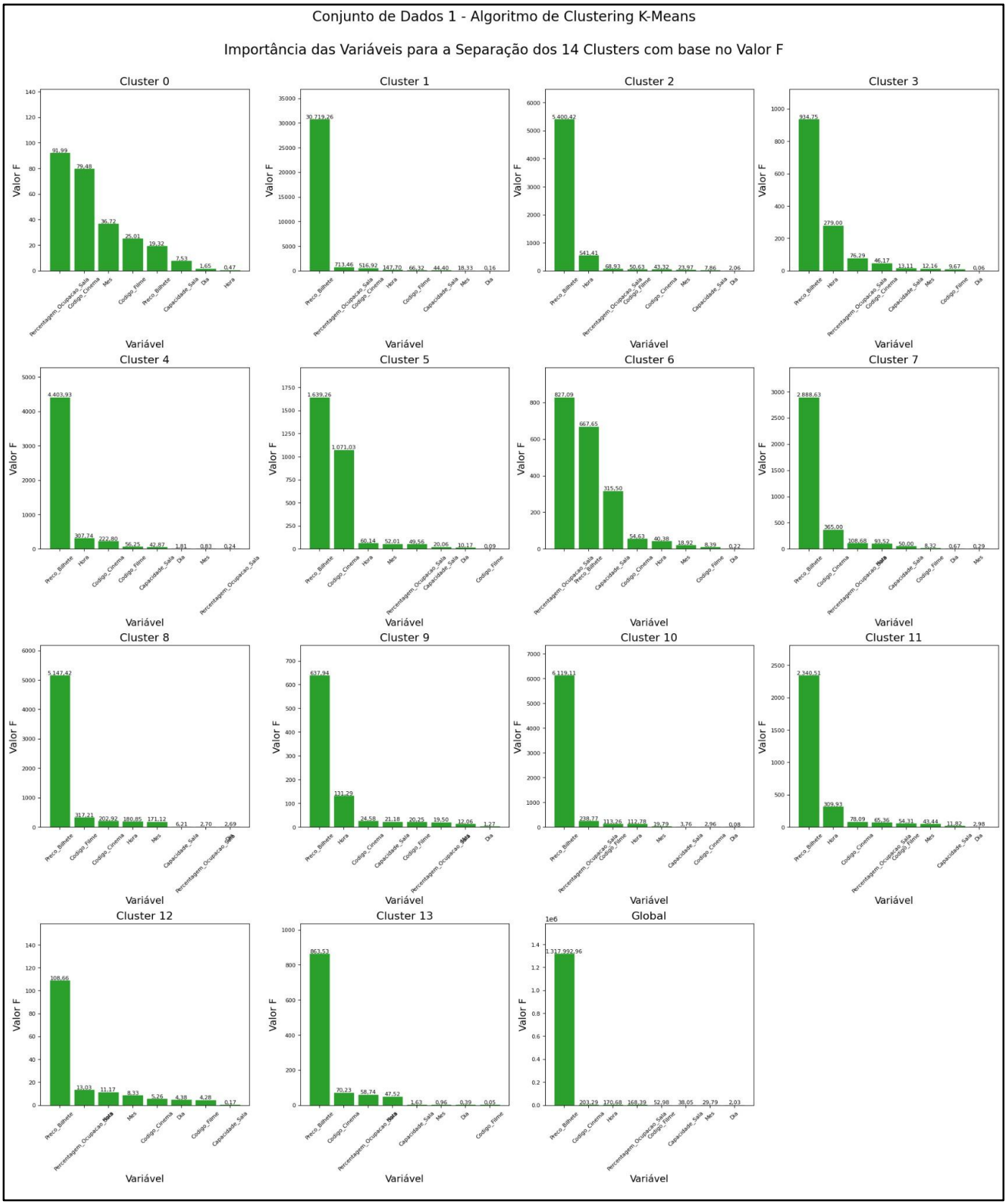

Figura d1.14 - Conjunto de Dados 1 - Algoritmo de Clustering K-Means - Importância das Variáveis para a Separação dos 14 Clusters com base no Valor F.

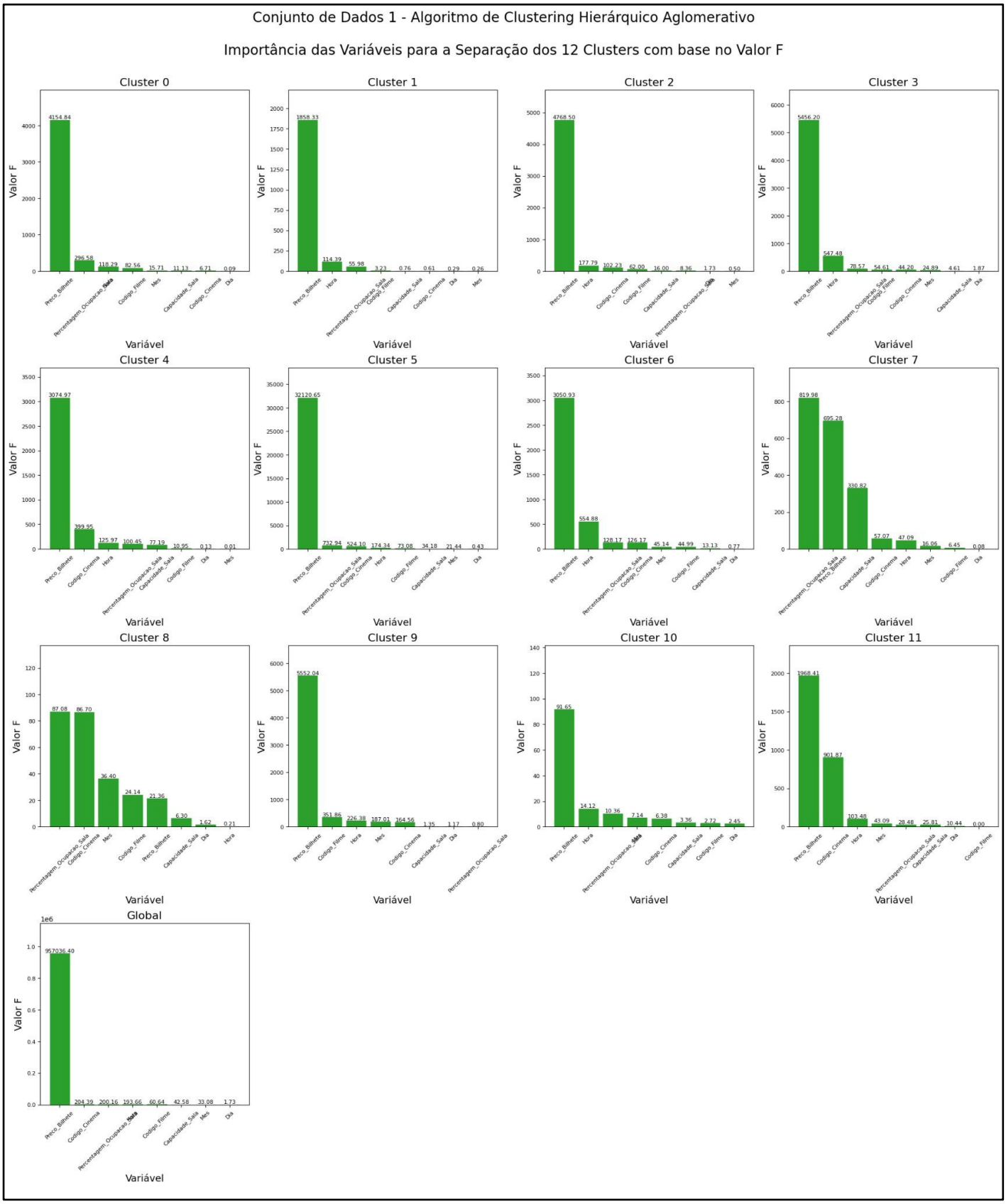

Figura d1.15 - Conjunto de Dados 1 - Algoritmo de Clustering Hierárquico Aglomerativo - Importância das Variáveis para a Separação dos 12 Clusters com base no Valor F.

Finalmente, aplicou-se o método de análise de perfil de clusters para os caracterizar individualmente, revelando padrões específicos em cada um deles. Neste método, em primeiro lugar, é calculada a média de cada variável por cluster, recorrendo ao conjunto de dados rotulado segundo os clusters identificados. De seguida, calcula-se a média global de cada variável, utilizando apenas o conjunto de dados original. Por último, é calculada a diferença entre as médias mencionadas, dividindo-se o resultado final pela média global. Assim, foi construída a Tabela d1.6, correspondente ao cálculo destas diferenças relativas para o K-Means. O principal objetivo deste método é caracterizar e interpretar os clusters formados, evidenciando os desvios que permitem compreender os perfis de cada um deles e identificar os padrões específicos que os distinguem do conjunto total. É de salientar que, inicialmente, o método foi aplicado considerando ambos os algoritmos; no entanto, dada a grande similitude dos resultados obtidos e de forma a evitar repetições, optou-se por apresentar apenas os que dizem respeito ao K-Means.

Tabela d1.6 - Algoritmo de Clustering K-Means - Diferença Relativa entre a Média da Variável por Cluster (14 Clusters) e a Média Global de cada Variável.

Conjunto de Dados 1 - Algoritmo de Clustering K-Means								
Diferença Relativa entre a Média de cada Variável por Cluster e a Média Global de cada Variável								
Cluster	Mes	Dia	Hora	Codigo_Filme	Codigo_Cinema	Preco_Bilhete	Capacidade_Sala	Percentagem_Ocupacao_Sala
0	-0.02	-0.01	0.00	0.00	-0.05	-0.02	-0.02	0.10
1	0.02	0.00	0.12	0.00	0.17	0.84	-0.09	0.43
2	-0.02	0.01	-0.20	0.00	0.05	-0.39	-0.03	-0.12
3	0.06	-0.01	0.52	0.00	0.17	0.61	0.15	0.45
4	-0.01	-0.02	-0.28	0.00	0.19	-0.64	-0.14	0.01
5	-0.03	0.02	0.07	0.00	-0.23	0.23	0.05	-0.10
6	0.02	0.00	0.05	0.00	-0.05	-0.14	0.20	-0.40
7	0.00	-0.01	-0.08	0.00	0.12	-0.26	-0.07	-0.14
8	0.07	-0.02	0.15	0.01	-0.12	0.47	-0.04	-0.03
9	0.04	0.02	0.25	0.00	0.09	0.36	0.14	0.16
10	-0.03	0.00	-0.12	0.00	-0.01	-0.52	-0.03	0.28
11	0.08	-0.04	0.42	0.01	0.17	0.73	0.11	0.32
12	-0.03	0.03	0.06	0.00	0.03	0.12	0.01	0.11
13	0.02	-0.02	-0.29	0.00	-0.28	-0.78	0.07	-0.53

2.7 Impacto do Pré-Processamento

Inicialmente, foi aplicada uma normalização às variáveis de forma a reduzir a sua escala, utilizando o método de normalização Z-Score. No entanto, ao calcular os valores do índice de silhueta global mediante as diferentes combinações de parâmetros usados para cada algoritmo de clustering, constatou-se que o valor máximo desse índice rondava apenas os 0.20, valor muito baixo. Após várias tentativas seguindo diferentes abordagens para a sua maximização (não remover valores atípicos, excluir variáveis temporais, etc.), verificou-se que, ao não normalizar os dados, este índice subiu dos 0.20 para os 0.77, valor muitíssimo mais alto. Pode considerar-se a hipótese de as escalas das variáveis refletirem já a relevância e a variabilidade intrínseca de cada uma delas, permitindo a formação de clusters bem definidos; a normalização poderia ter igualado a influência das variáveis, omitindo diferenças essenciais a uma segmentação correta dos dados. É uma hipótese igualmente válida considerar que a normalização atenuaria variações relevantes, diminuindo as diferenças naturais entre os clusters. Assim, optou-se por não efetuar a normalização de escala, sendo que todas as transformações anteriormente descritas na secção relativa ao data profiling e ao pré-processamento de dados foram realizadas e mantidas.

2.8 Avaliação Detalhada

Para o K-Means, analisando os valores do índice de silhueta global, verifica-se que, se se utilizarem apenas 2 ou 3 clusters, a

segmentação não consegue capturar as nuances existentes nos dados. À medida que se aumenta o número de clusters para 5, 6 e 7, o valor do índice de silhueta global sobe gradualmente, alcançando os 0.666 e até aos 0.690. Verifica-se uma subida mais acentuada com 8 clusters (0.763) e, apesar de pequenas flutuações no aumento para 9 e até aos 13 clusters, o pico máximo registado dá-se para 14 clusters (0.773). Para esta segmentação, o resultado mostra que é atingido um equilíbrio ótimo entre a coesão interna e a separação entre os clusters com 14 grupos, sendo os valores de silhueta superiores a 0.7 geralmente indicativos de clusters bem definidos. Uma hipótese plausível para este facto é que, à medida que o número de clusters vai aumentando, dá-se um agrupamento mais preciso de pontos com características semelhantes, melhorando a coesão interna. Mas, ao aumentar para 15 clusters, verifica-se uma queda, podendo indicar um fenómeno de overfitting na segmentação, isto é, grupos naturalmente homogéneos serem divididos de forma excessivamente detalhada. Com 14 clusters, é possível um alinhamento do algoritmo com a distribuição intrínseca dos dados; números superiores distorcerão essa relação.

Quanto ao Aglomerativo, a ligação simples registou de forma consistente valores de índice de silhueta global muito baixos (0.097 e -0.113), mostrando que não é adequada para o conjunto de dados em questão. As ligações média e completa apresentaram desempenhos intermédios, com a média a revelar uma tendência de melhoria contínua, atingindo 0.728 para 15 clusters. A ligação de ward destacou-se como a melhor escolha, já que apresentou sempre os valores mais elevados do índice de silhueta, atingindo um máximo de 0.762 para 12 clusters. Conclui-se deste resultado que a ligação de ward possibilitou uma segmentação mais coesa e bem separada. Utilizando 12 clusters e o método de ligação de ward, foram testadas as distâncias euclidiana, manhattan e cosseno. A distância euclidiana apresentou um índice de silhueta global de 0.762, tendo a distância manhattan obtido um valor ligeiramente inferior (0.740). Por sua vez, a distância cosseno teve um desempenho bastante insatisfatório (-0.122), depreendendo-se que não capta convenientemente a estrutura dos dados neste contexto. O método de ligação de ward permite minimizar a variância intra-cluster, adequando-se a distância euclidiana a este critério de otimização. Assim, a minimização dos desvios quadráticos favorece a formação de clusters mais compactos e bem separados, e contribui para um índice de silhueta mais elevado. Uma possibilidade a considerar é que os clusters presentes no conjunto de dados tenham uma forma sensivelmente esférica e uma distribuição homogénea, adequando-se a distância euclidiana a estas características.

Os valores mais elevados dos índices de silhueta globais para ambos os algoritmos indicam a presença de clusters bem definidos. A elevada proximidade destes 2 valores sugere que a estrutura de clustering dos dados é bastante estável. Para o K-Means, mesmo quando os clusters são bem definidos, o valor da inércia pode ser elevado quando não é efetuada uma normalização de escala das variáveis durante o pré-processamento de dados.

Para o Aglomerativo, o valor baixo do índice de Davis-Bouldin (0.135) aponta para clusters compactos e bem separados, com pouca sobreposição entre si. Os bons resultados obtidos para ambos os índices internos deste algoritmo permitem concluir que o mesmo apresenta um bom desempenho.

Quanto aos índices externos, o valor de pureza de 0.45 significa que, em média, apenas 45% dos registos de cada cluster correspondem à classe maioritária da variável-alvo (*Receita_Total_Discretizada*). Já o índice de rand de 0.02 indica uma correspondência muito baixa entre os clusters criados e as classes reais. Estes valores sugerem um fraco alinhamento entre os clusters obtidos por ambos os algoritmos e as classes reais da variável-alvo. Uma hipótese a considerar é que os dados com alta variabilidade interna dificultam a identificação de clusters coerentes em relação à variável-alvo, pelo que podem não apresentar uma estrutura natural que permita um forte alinhamento.

As soluções de clustering mais promissoras utilizando ambos os algoritmos demonstraram que os clusters formados se encontram bem separados, sem sobreposições significativas. Deste modo, as parametrizações escolhidas para cada tipo de algoritmo revelaram-se bastante eficazes, tanto para a formação como para a separação dos clusters. Cada uma das componentes principais 1 e 2 explicou 24% da variância total dos dados, indicando que ambas apresentaram uma contribuição equilibrada na representação de padrões existentes. É importante ter em atenção que não se representou 52% da variância, podendo ter sido omitidas informações relevantes na estrutura dos dados. No entanto, mesmo considerando apenas 48% da variância total, as soluções apresentadas levaram a bons resultados.

Para ambos os algoritmos, a maioria dos clusters revela uma alta consistência entre os respetivos centróides, medianas e medoides, com diferenças insignificantes entre essas medidas (na ordem dos 0.00 até 0.08 para o K-Means e 0.00 até 0.10 para o Aglomerativo), indicando distribuições simétricas, pelo que esses clusters são bem definidos e homogéneos. Além disso, apresentam uma baixa dispersão média, sugerindo uma boa coesão interna dos dados. No entanto, alguns clusters exibem discrepâncias significativas entre as suas medidas de tendência central. Por exemplo, o cluster 2 do K-Means apresenta um centróide (-0.18) bem afastado da sua mediana (0.10) e do seu medoide (0.06), evidenciando uma assimetria negativa com possível influência de valores atípicos. Estes últimos clusters têm uma alta dispersão média (com valores como 2.20 no mencionado cluster 2), mostrando heterogeneidade interna.

Apenas para efeitos comparativos, apresenta-se na Figura d1.16 uma solução de clustering com o algoritmo Aglomerativo, utilizando uma combinação de parâmetros de 10 clusters com uma ligação mínima, que levou a um índice de silhueta global

extremamente baixo (0.097). É possível observar uma nuvem ampla e contínua de pontos, sendo que as baixíssimas coesão interna e separação entre os clusters resultaram em diversas sobreposições, com centróides muito próximos entre si. Dada a grande diferença de resultados entre ambas as soluções utilizando aquele algoritmo, conclui-se da enorme importância de uma boa parametrização para a formação e separação dos clusters.

A análise dos valores F permitiu identificar as variáveis com maior influência na segmentação dos dados. Globalmente, para ambos os algoritmos, o valor F da variável *Preço_Bilhete* é extremamente elevado (1.317.992,96 para o K-Means e 957.036,26 para o Aglomerativo), indicando que ela tem, de longe, o maior impacto na separação dos clusters. Na quase totalidade destes últimos, considerados individualmente, a *Preco_Bilhete* apresenta os valores F mais altos, sugerindo que os grupos identificados por ambos os algoritmos diferem fortemente em relação ao preço dos bilhetes de cinema. É também possível observar que a variável *Percentagem_Ocupacao_Sala* se encontra entre as mais importantes. Para o K-Means, os clusters 6 e 10 apresentam valores F relativamente altos para esta variável (681,55 e 827,09, respetivamente); para o Aglomerativo, o cluster 7 apresenta um valor de 819,98. Isto sugere que os grupos identificados por ambos os algoritmos variam de forma bastante expressiva em relação à ocupação das salas de cinema. Globalmente, as variáveis *Codigo_Cinema* e *Hora* têm uma importância moderada, na ordem dos valores F de 200 para *Codigo_Cinema* e de 180 para *Hora* (considerando ambos os algoritmos). A variável *Codigo_Cinema* pode refletir características específicas dos estabelecimentos, tais como localização, infraestrutura e perfil do público, pelo que os cinemas podem ser agrupados com base em atributos comuns. A hora de exibição dos filmes indica, possivelmente, diferentes padrões de horário em que os filmes são mais ou menos populares. Embora a variável *Capacidade_Sala* apresente alguns valores F relevantes para a separação dos clusters (315,50 para o cluster 6 do K-Means e 330,82 para o 7 do Aglomerativo), os seus valores F são baixos na análise global (na ordem dos 40, para ambos os algoritmos). Deste modo, a variável tem influência em alguns clusters específicos, mas não é um fator determinante na segmentação. O mesmo acontece com a variável *Codigo_Filme*, que apresenta 2 valores F relevantes para a segmentação (317,21 para o cluster 8 do K-Means e 351,86 para o 9 do Aglomerativo); no entanto, os seus valores F globais são baixos (na ordem dos 50-60 para ambos os algoritmos). Os valores F para as variáveis *Mes* e *Dia* são consistentemente baixos na maioria dos clusters, indicando que aquelas contribuem pouco para a segmentação dos dados. Assim, as variações ao longo dos meses e dos dias não constituem um fator determinante para diferenciar os grupos. Desta maneira, as separações dos clusters realizadas por ambos os algoritmos são fortemente influenciadas pelas variações de *Preço_Bilhete*, seguido das variáveis *Percentagem_Ocupacao_Sala* e *Hora*. No entanto, existe uma diferença muito pronunciada entre a importância de *Preço_Bilhete* e as das restantes variáveis. Uma hipótese para esta situação poderá residir na escala dos valores de *Preço_Bilhete*, consideravelmente mais elevados do que os das outras variáveis. Assim, esta variável revela-se a mais importante para o clustering.

Por fim, através do cálculo das diferenças relativas entre a média de cada variável por cluster e a média global de cada variável, foram identificados padrões específicos entre os clusters, agrupando-se estes em 5 conjuntos. Para cada conjunto, foi, de seguida, colocada uma hipótese que pudesse explicar os desvios detetados, traçando possíveis perfis referentes a cada conjunto. Na formação destes, foram tidas em conta apenas as variáveis mais importantes para a separação dos clusters, *Preco_Bilhete* e *Percentagem_Ocupacao_Sala*, as quais apresentaram desvios significativos.

Para o K-Means, os conjuntos de clusters identificados e respetivos perfis são os seguintes:

1. Os clusters 1, 3 e 11 apresentaram desvios relativos a preços de bilhetes consideravelmente acima da média (0.84, 0.61 e 0.73, respetivamente) e desvios moderadamente positivos de percentagens de ocupação da sala (0.43, 0.45 e 0.32, respetivamente). A hipótese colocada é que o público estará disposto a pagar preços de bilhetes mais altos, mantendo uma boa adesão aos filmes em questão, pelo que um possível perfil para este conjunto poderá ser *Sessões Premium com Grande Procura*.

2. Os clusters 2, 6, 7 e 13 mostraram desvios negativos tanto para preços de bilhetes (-0.39, -0.14, -0.26 e -0.78, respetivamente), como para percentagens de ocupação da sala (-0.12, -0.40, -0.14 e -0.53, respetivamente). Uma possível justificação é que os preços mais baixos não serão suficientes para gerar uma procura significativa; assim, um perfil para este conjunto poderá ser *Sessões Económicas com Baixa Procura*.

3. Os clusters 4 e 10 mostraram desvios de preço muito inferiores à média (-0.64 e -0.52, respetivamente) e desvios distintos para a percentagem de ocupação da sala (0.01 e 0.28, respetivamente). Uma hipótese a considerar é que estas sessões evidenciam uma estratégia de preços baixos, no entanto, isso não garante uma boa adesão dos espetadores. A diferença entre estes clusters indicará, então, que, para além do preço, outros fatores influenciam a resposta do público. Um perfil para este conjunto poderá ser *Sessões Económicas e Procura Irregular*.

4. Os clusters 5, 8 e 9 apresentaram desvios de preços moderadamente positivos (0.23, 0.47 e 0.36, respetivamente),

sendo que a percentagem de ocupação da sala é muito próxima de 0 (-0.10, -0.03 e 0.16). Estas sessões de cinema terão uma política de preços um pouco mais elevada, mas a estratégia não altera substancialmente a percentagem de ocupação da sala, que permanece dentro da média. A este conjunto poderá corresponder o perfil *Sessões com Preços Moderadamente Elevados e Procura Normal*.

5. Os clusters 0 e 12 evidenciaram desvios de preços muito próximos de 0 (-0.02 e 0.12, respetivamente), com desvios de percentagem da ocupação da sala também muito próximos de 0 (0.10 e 0.11, respetivamente). Neste caso, os clusters corresponderão a sessões equilibradas, onde a oferta e a procura se mantêm em níveis consistentes. Um possível perfil para este conjunto é *Sessões a Preço Normal e Ocupação Moderada*.

2.9 Conclusões

Foi aplicada uma Análise de Componentes Principais (PCA), de forma a extrair as duas componentes principais de maior relevância. Quanto aos índices internos, o conjunto de dados apresenta um índice de silhueta global de aproximadamente 0.77 para ambos os algoritmos, mostrando clusters bem definidos, com uma boa coesão interna e separação clara entre si. No entanto, a elevada inércia observada no K-Means deve-se, provavelmente, à não normalização das variáveis, o que sugere que a escala destas influencia substancialmente a formação dos clusters.

Em termos de índices externos, o conjunto apresenta uma pureza de 0.45 e um índice de rand de 0.02 para ambos os algoritmos; assim, apesar de a estrutura interna dos clusters ser boa, o alinhamento com a variável-alvo é fraco, podendo ser explicado pela alta variabilidade interna dos dados.

Portanto, a robustez interna dos clusters não se traduz num bom alinhamento com as classes reais, possivelmente devido à influência de variáveis não normalizadas e à heterogeneidade dos dados.

As representações bidimensionais utilizando as Componentes Principais 1 e 2 mostraram uma boa formação e separação dos clusters.

3. ANÁLISE DE VALORES ATÍPICOS

Na fase seguinte do projeto, foi efetuada uma análise multivariada não-supervisionada de valores atípicos.

3.1 Seleção de Variáveis Relevantes

Após a fase de clustering, tendo em conta as 9 variáveis restantes presentes neste conjunto de dados, foram selecionadas as 4 consideradas mais relevantes para a análise de valores atípicos: *Preco_Bilhete*, *Percentagem_Ocupacao_Sala*, *Hora* e *Capacidade_Sala*. As 3 primeiras variáveis foram selecionadas por terem sido as mais importantes na separação dos clusters. Já a variável *Capacidade_Sala* foi escolhida porque fornece o contexto necessário à interpretaçãoda variável *Percentagem_Ocupacao_Sala*: uma salacom grande capacidade e baixa ocupação pode sinalizaruma sessão atípica, do mesmo modo que uma sala pequena com ocupação total pode igualmente apresentar um comportamento inesperado. Assim sendo, esta variável ajuda a identificar discrepâncias relevantes.

3.2 Modelo e Parâmetros

Para a deteção dos valores atípicos, optou-se por utilizar o modelo Local Outlier Factor (LOF), cujos parâmetros principais incluem o número de vizinhos (n) e a medida de distância. Foram testadas várias combinações destes parâmetros, considerando um intervalo de valores de 15 a 55 vizinhos, e as distâncias euclidiana, manhattan e do cosseno, sendo calculada a respetiva pontuação média dos valores atípicos (Tabela d1.6). Para outro parâmetro, a contaminação, foi assumido um limiar automático. Deste modo, selecionou-se a melhor combinação dos parâmetros, que minimizou a pontuação

média dos valores atípicos (1.243), tendo sido utilizados 25 vizinhos e a distância do cosseno.

Tabela d1.6 - Modelo de Deteção de Valores Atípicos Local Outlier Factor - Pontuação Média dos Valores Atípicos em função dos parâmetros Número de Vizinhos (n) e Medida de Distância.

Conjunto de Dados 1 - Deteção de Valores Atípicos		
Modelo Local Outlier Factor - Parâmetros a Testar		Pontuação Média dos Valores Atípicos
Número de Vizinhos (n)	Medida de Distância	
15	Euclidiana	4.002
	Manhattan	4.467
	Cosseno	1.252
20	Euclidiana	4.195
	Manhattan	4.573
	Cosseno	1.245
25	Euclidiana	4.253
	Manhattan	4.594
	Cosseno	1.243
30	Euclidiana	4.041
	Manhattan	4.397
	Cosseno	1.270
35	Euclidiana	3.651
	Manhattan	3.977
	Cosseno	1.322
40	Euclidiana	3.294
	Manhattan	3.554
	Cosseno	1.363
45	Euclidiana	1.547
	Manhattan	1.631
	Cosseno	1.414
50	Euclidiana	1.556
	Manhattan	1.639
	Cosseno	1.491
55	Euclidiana	1.568
	Manhattan	1.653
	Cosseno	1.618

3.3 Visualização

Utilizando o modelo Local Outlier Factor com a melhor combinação de parâmetros previamente definida, foram construídos 6 gráficos de dispersão bivariados referentes às 4 variáveis mais relevantes (Figuras d1.17 a d1.22).

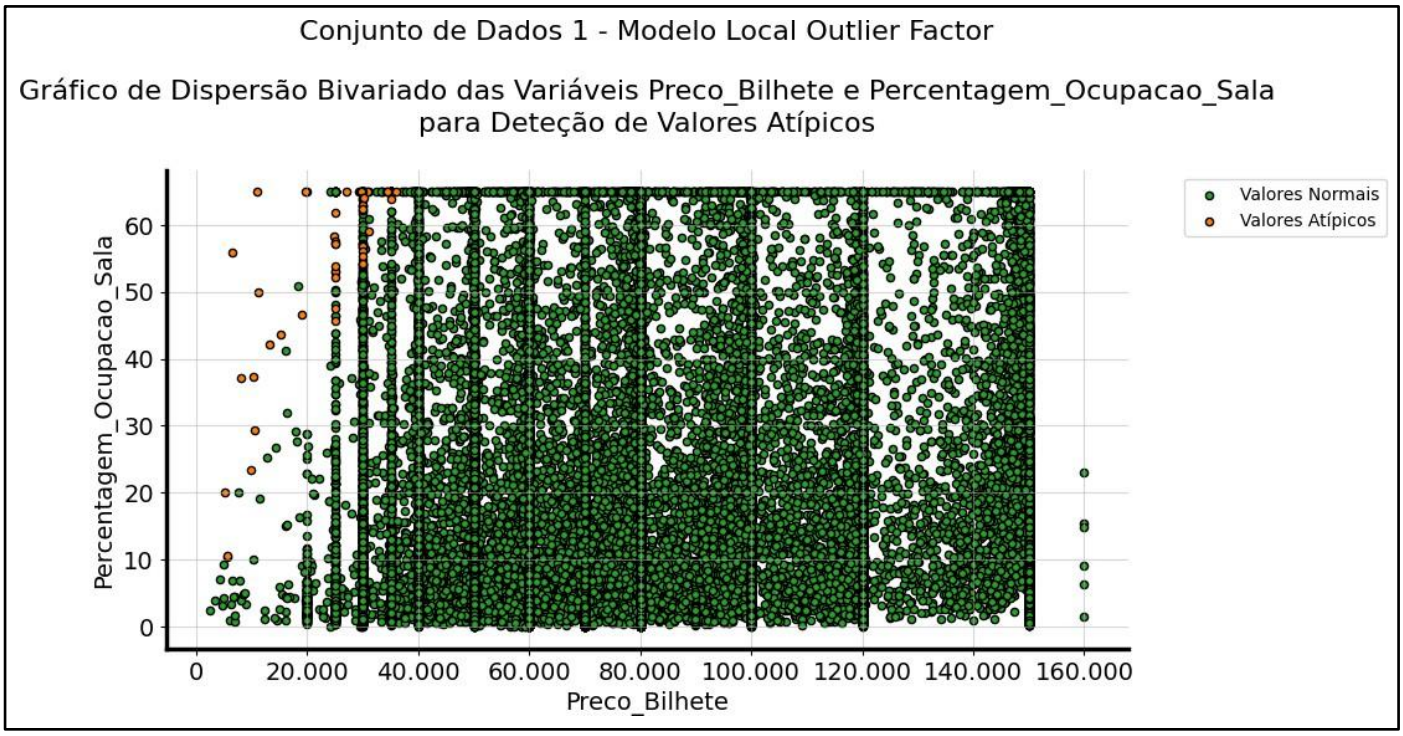

Figura d1.17 - Modelo Local Outlier Factor - Gráfico de Dispersão Bivariado das Variáveis Preco_Bilhete e Percentagem_Ocupacao_Sala para Deteção de Valores Atípicos.

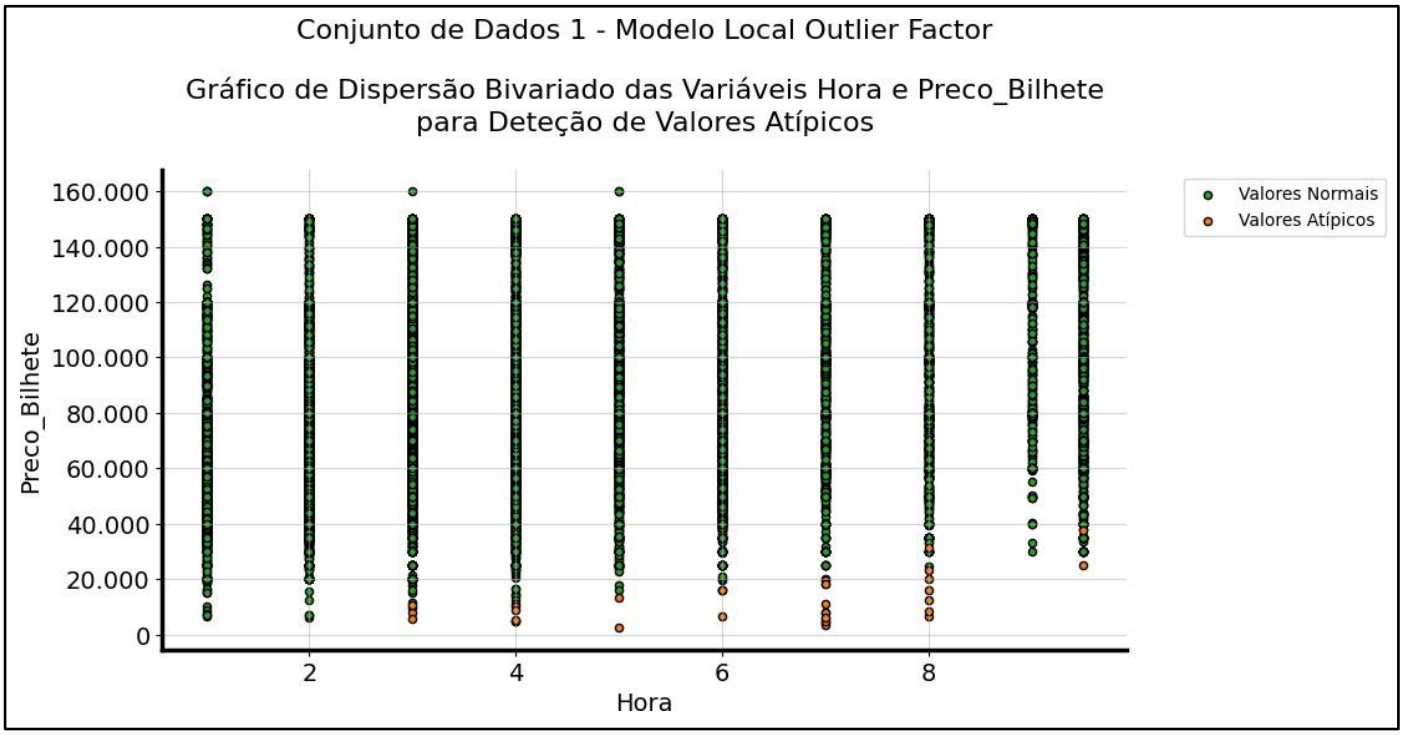

Figura d1.18 - Modelo Local Outlier Factor - Gráfico de Dispersão Bivariado das Variáveis Hora e Preco_Bilhete para Deteção de Valores Atípicos.

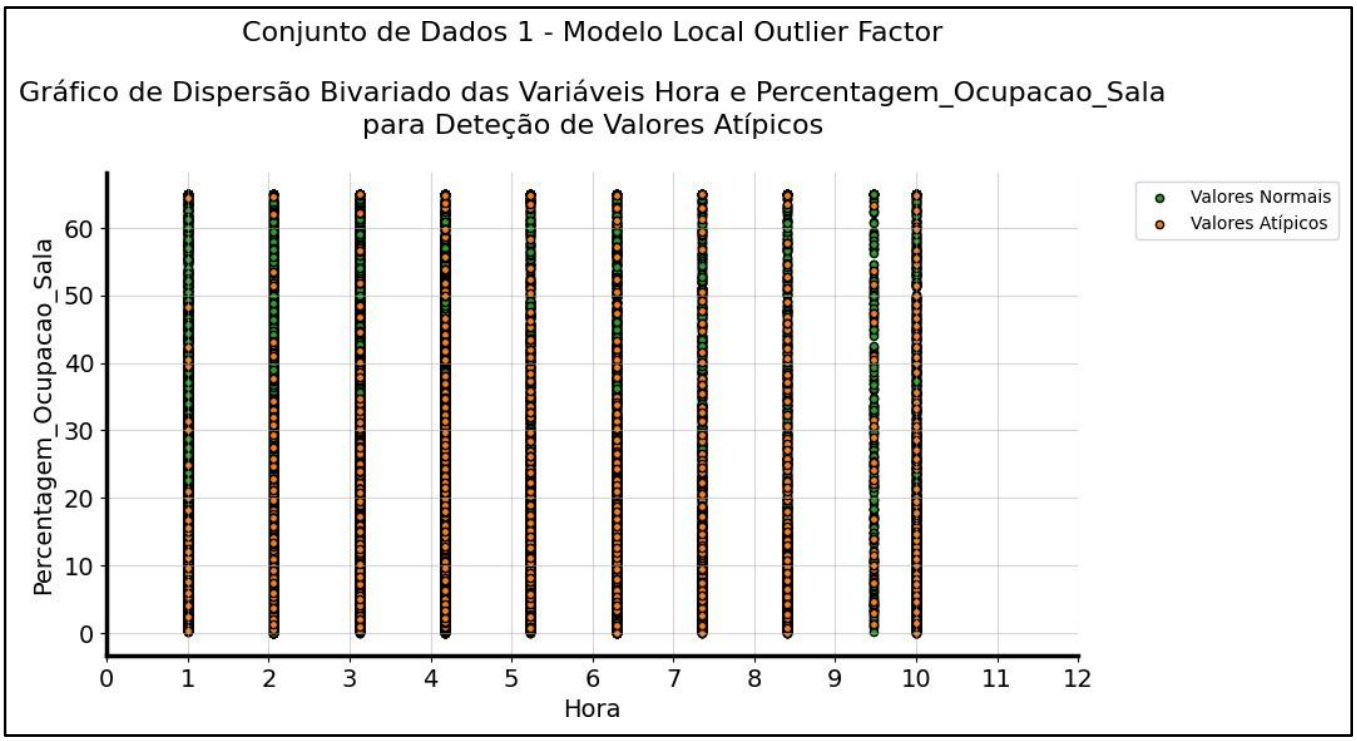

Figura d1.19 - Modelo Local Outlier Factor - Gráfico de Dispersão Bivariado das Variáveis Hora e Percentagem_Ocupacao_Sala para Deteção de Valores Atípicos.

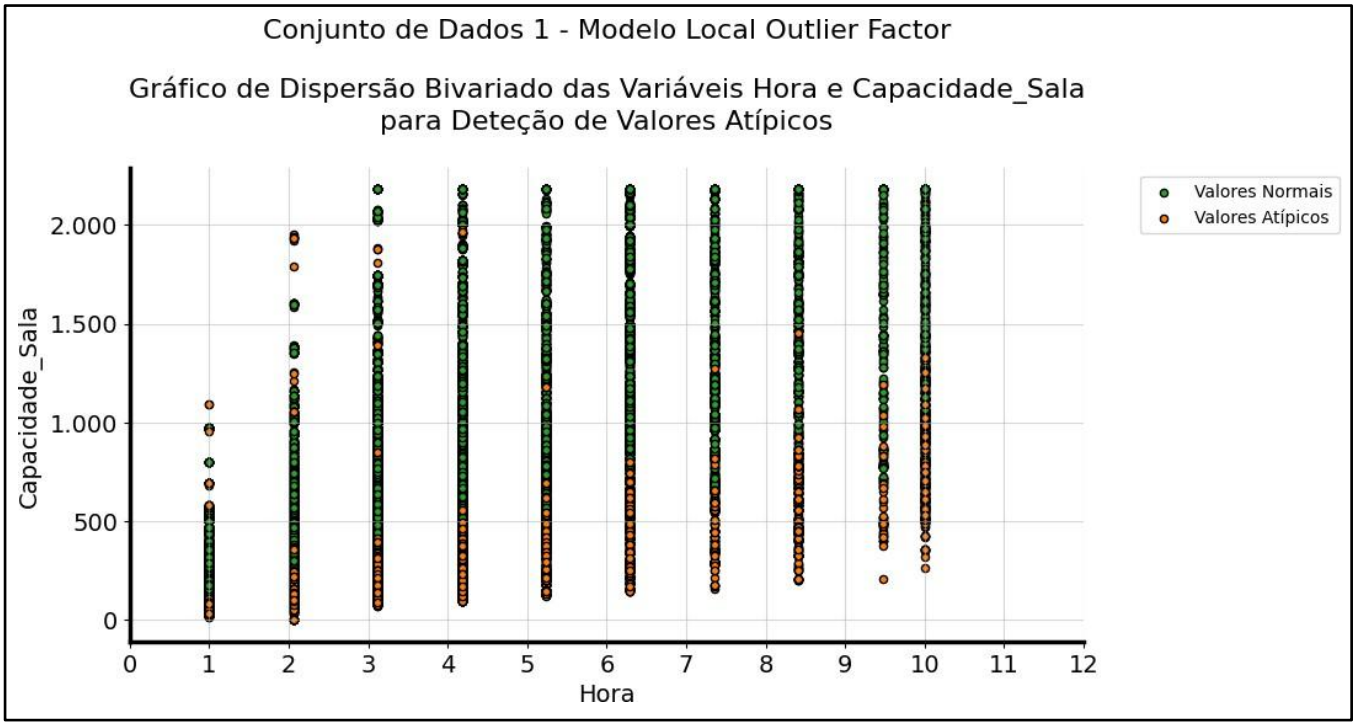

Figura d1.20 - Modelo Local Outlier Factor - Gráfico de Dispersão Bivariado das Variáveis Hora e Capacidade_Sala para Deteção de Valores Atípicos.

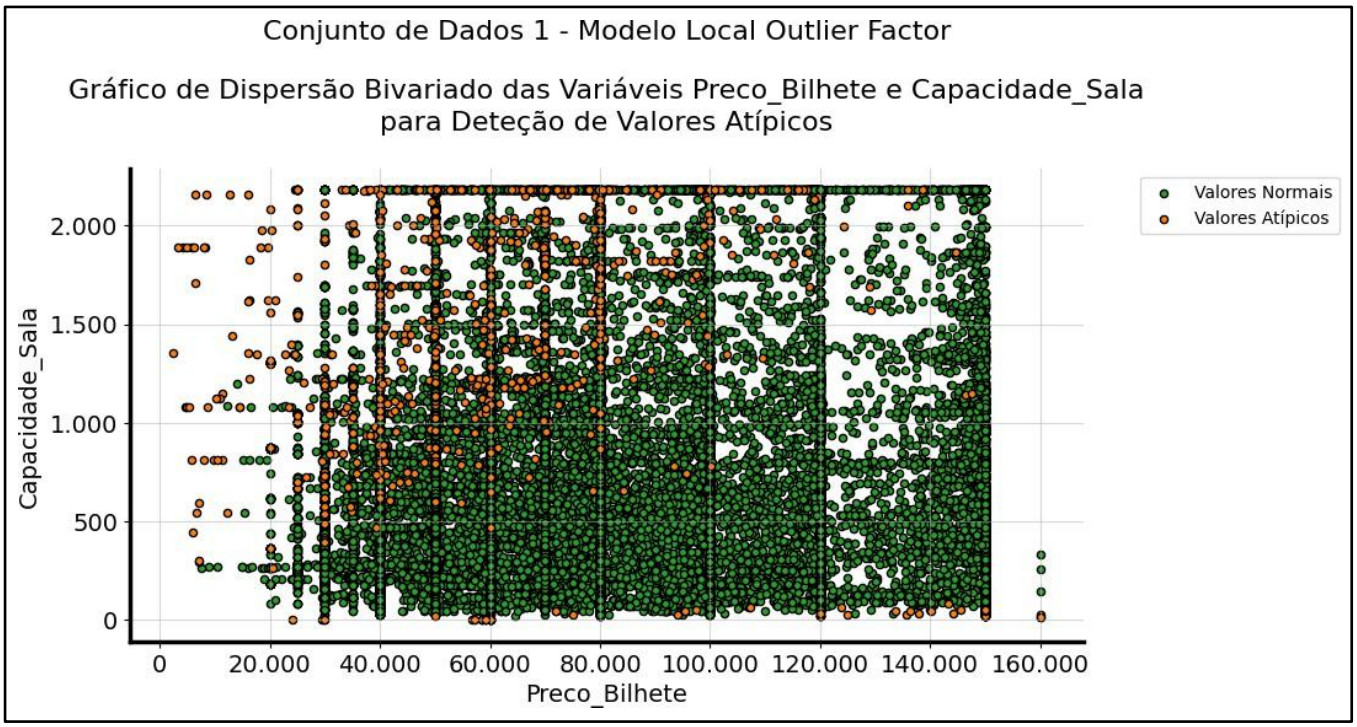

Figura d1.21 - Modelo Local Outlier Factor - Gráfico de Dispersão Bivariado das Variáveis Preco_Bilhete e Capacidade_Sala para Deteção de Valores Atípicos.

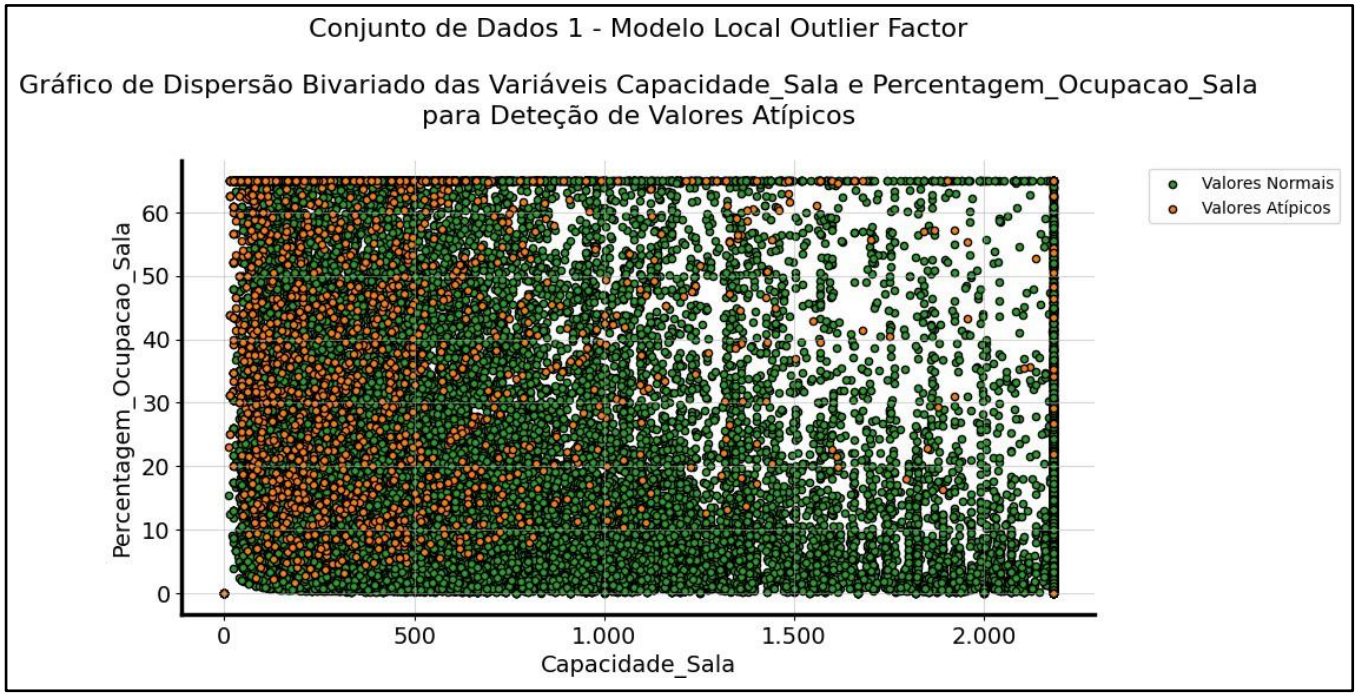

Figura d1.22 - Modelo Local Outlier Factor - Gráfico de Dispersão Bivariado das Variáveis Capacidade_Sala e Percentagem_Ocupacao_Sala para Deteção de Valores Atípicos.

3.4 Pontuação dos Valores Atípicos

Utilizando novamente o modelo Local Outlier Factor, com a melhor combinação de parâmetros, foram calculadas as pontuações dos valores atípicos. Relativamente à sua distribuição, encontram-se representados na Figura d1.23 um histograma e na Figura d1.24 um diagrama de caixa e bigodes.

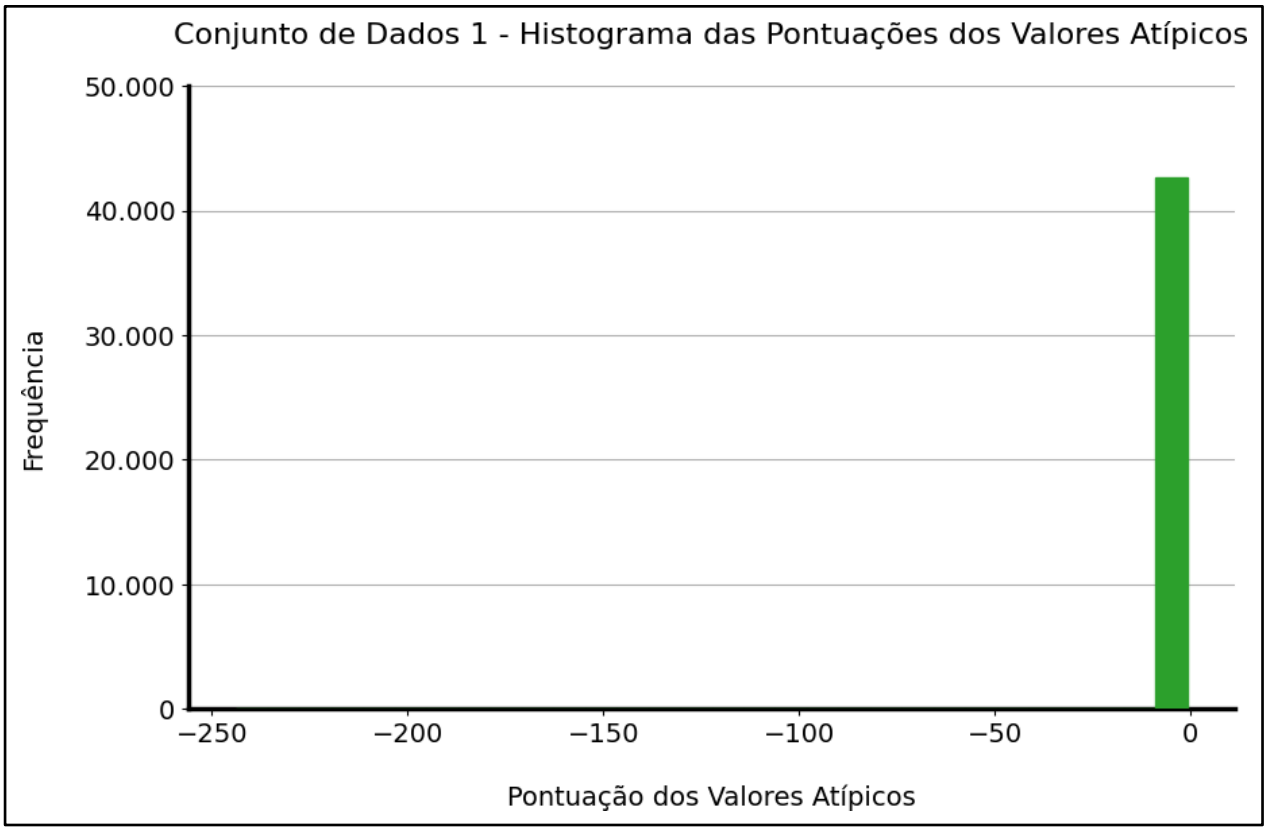

Figura d1.23 - Histograma das Pontuações dos Valores Atípicos.

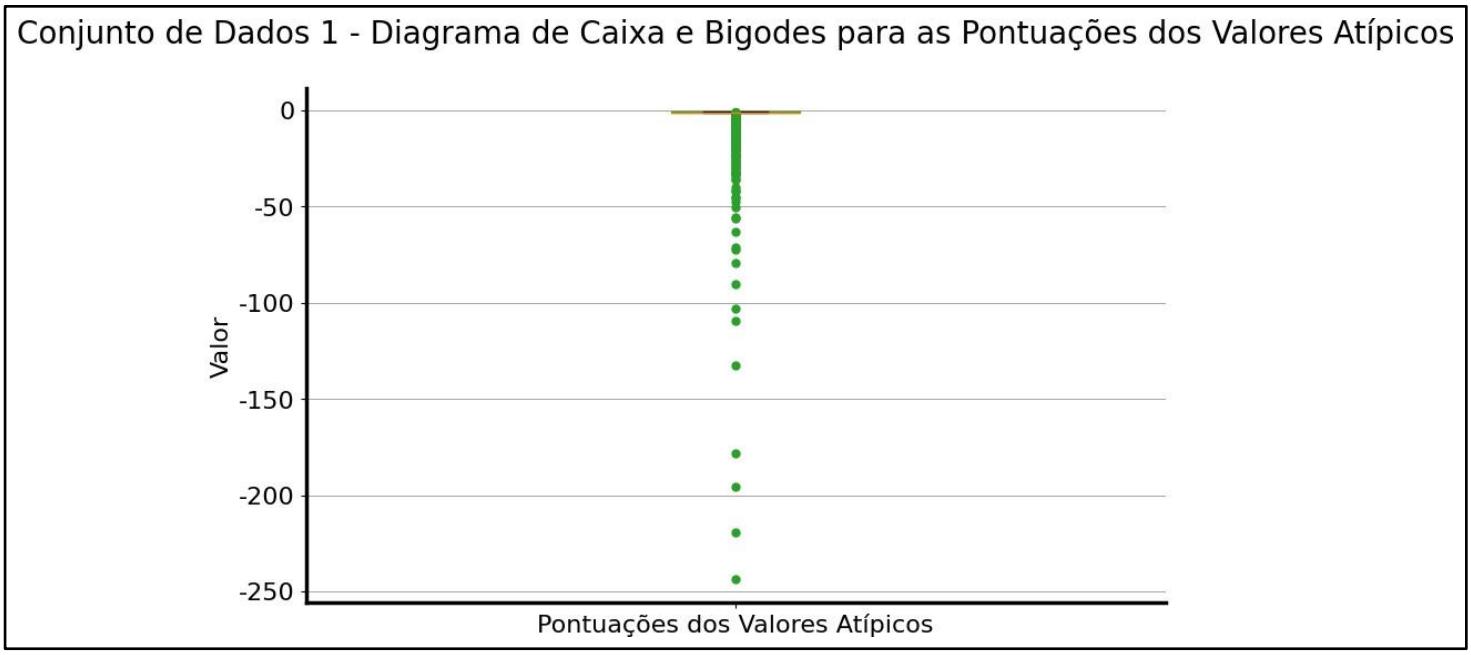

Figura d1.24 - Diagrama de Caixa e Bigodes para as Pontuações dos Valores Atípicos.

Por fim, de forma a distinguir os verdadeiros valores atípicos dos potenciais, foi novamente utilizado o método do Intervalo Interquartil (com um fator de 1.5), desta vez sobre o conjunto das pontuações dessas anomalias. Assim, verificou-se que a percentagem de verdadeiros valores atípicos rondava os 9.35%.

3.5 Avaliação Detalhada

No que se refere à minimização da pontuação média dos valores atípicos graças à melhor combinação dos parâmetros utilizados, verifica-seque, à medida que o valor do número de vizinhos (n) aumenta, as pontuações de anomalia para as distâncias euclidiana e manhattan tendem a diminuir, enquanto que para a distância cosseno vão aumentando. Observa-se, por outro lado, que com a medida de distância do cosseno as pontuações de anomalia são consistentemente mais baixas do que quando se utilizam as distâncias euclidiana e manhattan. Uma possível justificação é que a métrica de distância do cosseno avalia a semelhança entre os vetores com base na orientação e não na magnitude, sendo menos sensível às variações de escala dos dados, o que faz com que o algoritmo capte com mais exatidão a sua estrutura angular. Em dados dispersos, além disso, a comparação baseada em ângulos pode representar melhor as relações intrínsecas entre os pontos, o que torna mais fácil a identificação de valores atípicos. A pontuação de anomalia mais baixa foi de 1.243, registada para a combinação número de vizinhos igual a 25 e medida de distância de cosseno, mostrando que essa configuração foi a mais eficaz na deteção de valores atípicos neste conjunto de dados.

A análise dos gráficos de dispersão bivariados permitiram retirar as seguintes conclusões:

1. Par de Variáveis *Preco_Bilhete* e *Percentagem_Ocupacao_Sala*: O modelo Local Outlier Factor identificou apenas 66 valores atípicos num total de 42.270 registos, indicando que a relação entre o preço do bilhete e a percentagem de ocupação da sala segue um padrão muito consistente na grande maioria dos casos. Uma possível explicação para os pouquíssimos valores atípicos nas faixas de preço mais baixas (até ao valor de 40.000 de uma unidade monetária não especificada no website de origem do conjunto de dados) pode residir em promoções pontuais, como, por exemplo, bilhetes vendidos a preços reduzidos para determinadas sessões ou dias da semana, ou então para públicos específicos (crianças, idosos, estudantes, etc.). Além disso, pode haver casos raros de eventos gratuitos, identificados como anómalos. Em todos estes casos, não se verifica uma variação significativa relativamente à percentagem de ocupação da sala.

2. Par de Variáveis *Hora* e *Preco_Bilhete*: Foram identificados apenas 36 valores atípicos, representando uma fração muito baixa, isto é, os dados correspondem na quase totalidade a valores normais. Maioritariamente, os valores atípicos são referentes a preços de bilhetes mais reduzidos, distribuídos quase uniformemente pelas diversas horas de sessão. Estas anomalias podem derivar da realização de sessões especiais ou pré-estreias gratuitas, com preços eventualmente simbólicos devido a parcerias com festivais de cinema, ou, então, a erros de introdução dos dados (por exemplo, bilhetes inseridos incorretamente com preços nulos). Outra possibilidade reside no facto de os cinemas oferecerem bilhetes subsidiados aos seus funcionários ou a clientes de empresas parceiras. Estas situações ocorrem independentemente dos diversos horários.

3. Par de Variáveis *Hora* e *Percentagem_Ocupacao_Sala*: Neste par, foram identificados 3.736 valores atípicos, correspondendo a 9% do número total de registos, uma percentagem considerável. Estas anomalias ocorrem, principalmente, até 50% da ocupação da sala de cinema, com uma distribuição bastante homogénea para os diversos horários das sessões. Podem decorrer tanto de variações sazonais – períodos de férias ou idas à praia durante o verão, por exemplo – como da realização de eventos concorrentes - o caso dos jogos de futebol - os quais podem levar a taxas de ocupação consistentemente inferiores em todos os horários. Além disso, questões operacionais, tais como restrições temporárias de capacidade devido a operações de manutenção e segurança das salas, ou reconfiguração dos espaços, contribuem decerto para que a ocupação registada seja bastante inferior à capacidade das salas, resultando em percentagens abaixo de 50% em todas as faixas horárias.

4. Par de Variáveis *Hora* e *Capacidade_Sala*: Neste par, foram identificados 1.522 valores atípicos (4% do número total de registos). Na grande maioria dos casos, observa-se que eles estão associados a salas de cinema de menor capacidade, até 1.000 lugares, distribuídos sem nenhuma tendência clara para os diversos horários de exibição. Este fenómeno pode decorrer de situações particulares, como cinemas muito pequenos ou sessões especiais, que não seguem o padrão de ocupação e disponibilidade das restantes salas. Estes espaços poderão ter sido concebidos para oferecer experiências diferenciadas, independentemente dos horários de exibição. Salas VIP costumam ter uma lotação intencionalmente inferior para garantir exclusividade e conforto, privilegiando a qualidade da experiência cinematográfica em detrimento do número de lugares, situação que se verifica em qualquer horário. Outra possibilidade prende-se com a existência de cinemas históricos, que mantêm muitas das vezes o seu *design* original, e que, devido a limitações estruturais, apresentam capacidades reduzidas. Estes estabelecimentos, quando comparados com cinemas de maior dimensão, podem evidenciar valores atípicos.

5. Par de Variáveis *Preco_Bilhete* e *Capacidade_Sala*: Tal como no par de variáveis *Hora* e *Preco_Bilhete*, foi identificada uma pequena quantidade de valores atípicos (desta vez, 1.066, correspondentes a 2% do conjunto de dados), referentes a preços de bilhete mais baixos, na faixa dos 20.000 até aos 80.000, e a partir de uma capacidade de sala de cinema de 500 lugares. Uma vez mais, estes valores poderão resultar de promoções ou sessões especiais.

6. Par de Variáveis *Capacidade_Sala* e *Percentagem_Ocupacao_Sala*: No respetivo gráfico, existem 1.939 valores atípicos, correspondendo a 5% de registos do conjunto de dados total. Estes valores encontram-se maioritariamente concentrados nas salas de cinema de baixa capacidade, de até, aproximadamente, 1.000 lugares. No entanto, relativamente à variável *Percentagem_Ocupacao_Sala*, os valores atípicos aparecem disseminados de forma relativamente uniforme, sem nenhuma tendência associada. Uma hipótese para este fenómeno é que, em determinadas alturas, as salas de baixa capacidade podem ser reservadas para pré-estreias, sessões especiais, ou para a exibição de filmes de menor audiência, criando padrões de ocupação inconsistentes, ou seja, a períodos de elevada ocupação seguem-se períodos de subutilização da sala. Por outro lado, o número de pessoas necessário para alterar a percentagem de ocupação da sala de forma significativa é mais reduzido, pelo que pequenas flutuações podem ser interpretadas como anómalas pelo modelo Local Outlier Factor.

Observando o histograma e o diagrama de caixa e bigodes das pontuações dos valores atípicos, nota-se que a característica mais relevante é a elevada concentração dessas pontuações muito próxima do valor 0. Ela indica que a maior parte dos dados foi classificada como normal pelo modelo Local Outlier Factor, sugerindo uma distribuição predominantemente homogénea em termos do comportamento das variáveis analisadas. Mesmo assim, existem pontos situados muito abaixo do limite inferior do diagrama, mostrando a presença de valores extremos bastante negativos, que se estendem até aproximadamente -250. Este comportamento revela uma distribuição assimétrica, com uma cauda longa em direção aos valores negativos, que o histograma evidencia (apesar de não ser claramente visível no mesmo, dada a existência de pouquíssimos valores para esta faixa). A

extensão revela a existência de vários pontos efetivamente classificados como valores atípicos, mostrando pontuações muito diferentes das do grupo. No modelo Local Outlier Factor, os valores extremos indicam uma maior probabilidade de anormalidade, uma vez que os respetivos pontos se afastam das regiões de alta densidade do espaço de variáveis. O resultado obtido foi expectável, já que se previa que a maioria das observações fosse considerada normal, com apenas alguns valores a serem identificados como afastados do padrão geral.

Ao aplicaro Intervalo Interquartil ao conjunto das pontuações dosvalores atípicos, verificou-seque o número real de anomalias era de 3.993 (9.35% do número total de registos), o que corresponde, muito provavelmente, à quantidade de valores atípicos que o truncamento aos limites inferior e superior não conseguiu remover na fase do pré -processamento de dados, sendo posteriormente identificados nos gráficos de dispersão bivariados.

3.6 Conclusões

Foi utilizado o modelo Local Outlier Factor para a deteção dos valores atípicos. De forma a minimizar o valor das pontuações de anomalias, foi selecionada a combinação de 25 vizinhos e a medida de distância de cosseno em ambos os conjuntos de dados. A seleção desta medida para identificação de valores atípicos foi expectável, uma vez que é menos sensível a variações de escala, apresentando uma eficácia mais alta para dados dispersos.

Foi possível observar que, à medida que o número de vizinhos aumenta, as pontuações de anomalia associadas às distâncias euclidiana e manhattan tendem a diminuir, enquanto a pontuação relativa à distância do cosseno aumenta, mantendo-se consistentemente mais baixa do que as restantes.

Verificou-se que os dados apresentavam aproximadamente 9.35% de verdadeiros valores atípicos, provavelmente decorrentes de uma minoria de valores não removidos durante o pré-processamento de dados. Os valores do conjunto de dados apontam para a existência de uma quantidade significativa de anomalias, tendo sido apresentadas hipóteses para a sua existência nas variáveis mais relevantes durante a análise dos gráficos bivariados.

APÊNDICE

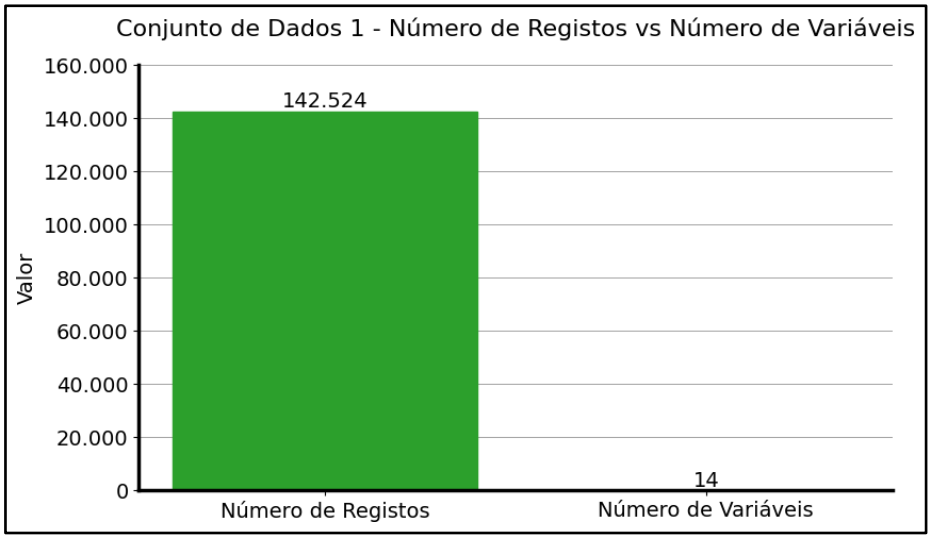

Figura d1.1 - Número de Registos vs Número de Variáveis.

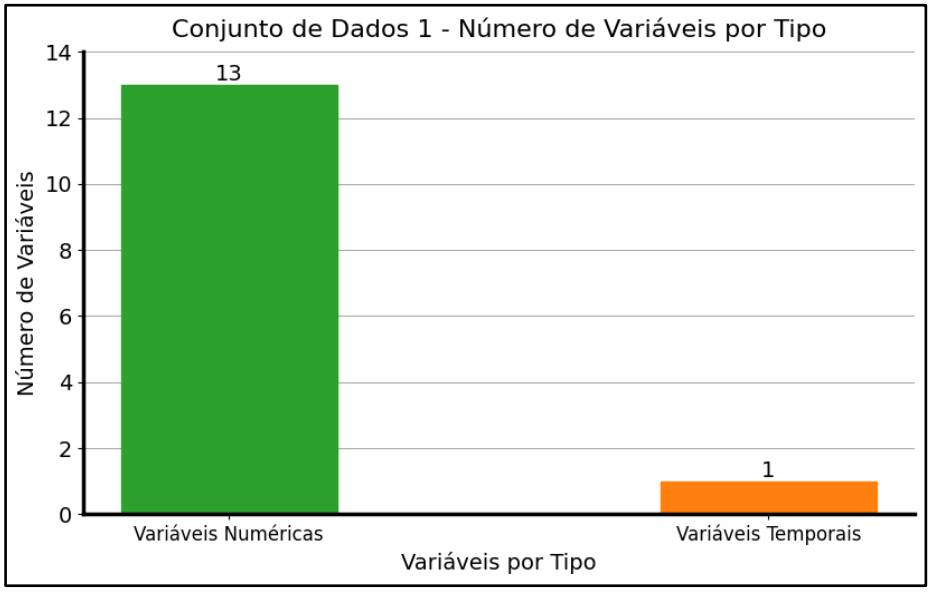

Figura d1.2 - Número de Variáveis por Tipo.

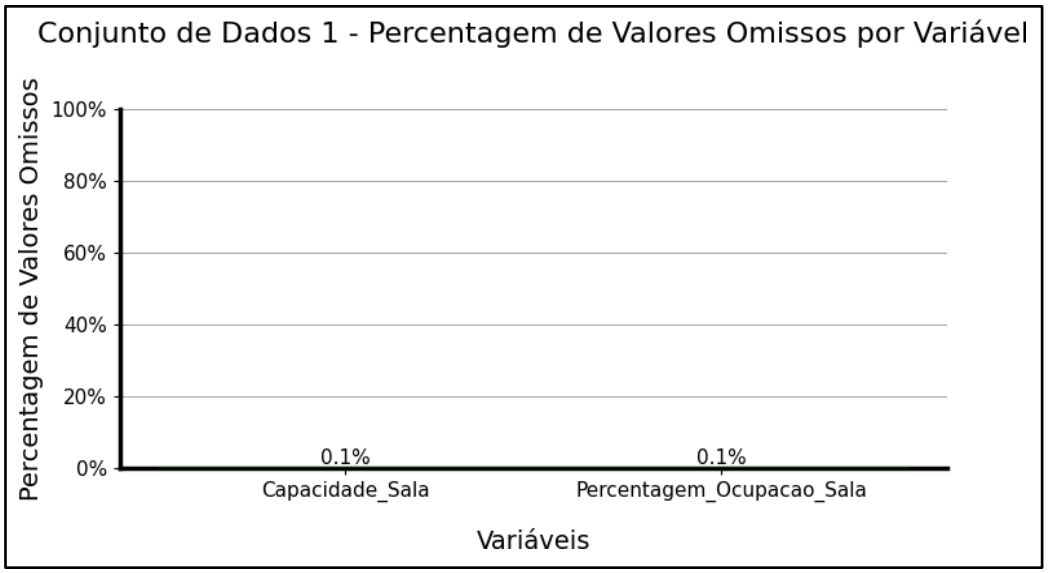

Figura d1.3 - Percentagem de Valores Omissos por Variável.

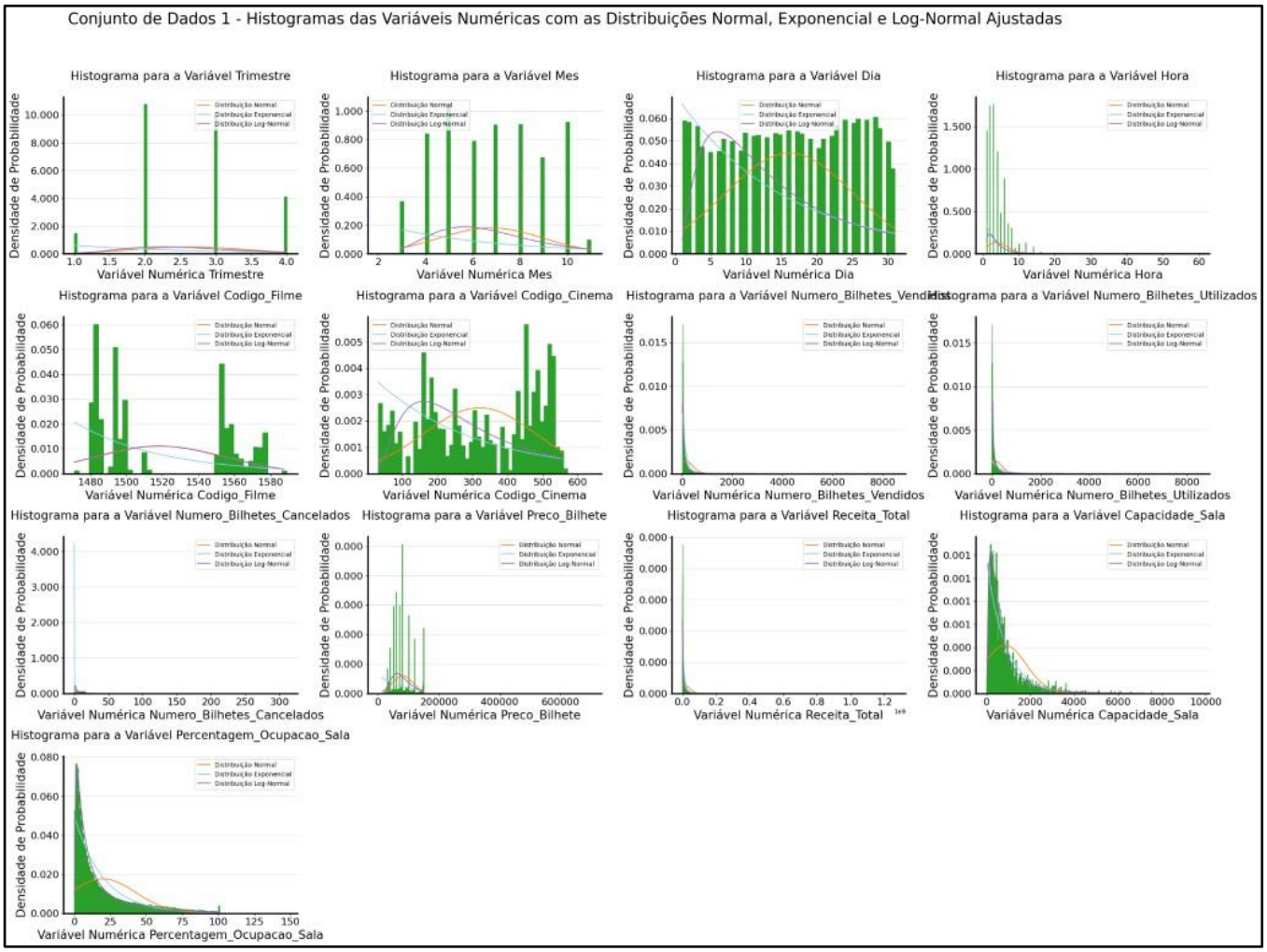

Figura d1.4 - Histogramas das Variáveis Numéricas com as Distribuições Normal, Exponencial e Log- Normal Ajustadas.

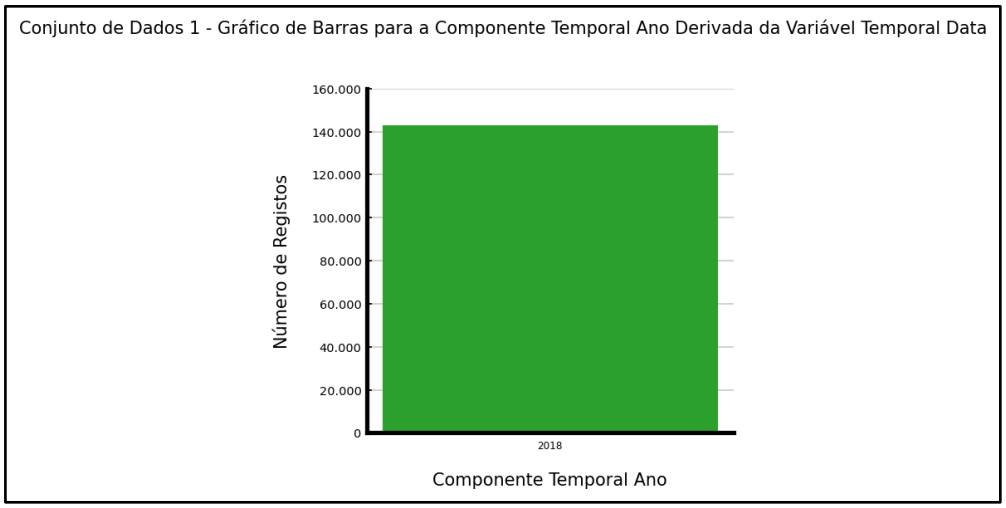

Figura d1.5 - Gráfico de Barras para a Componente Temporal *Ano* derivada da Variável Temporal *Data*.

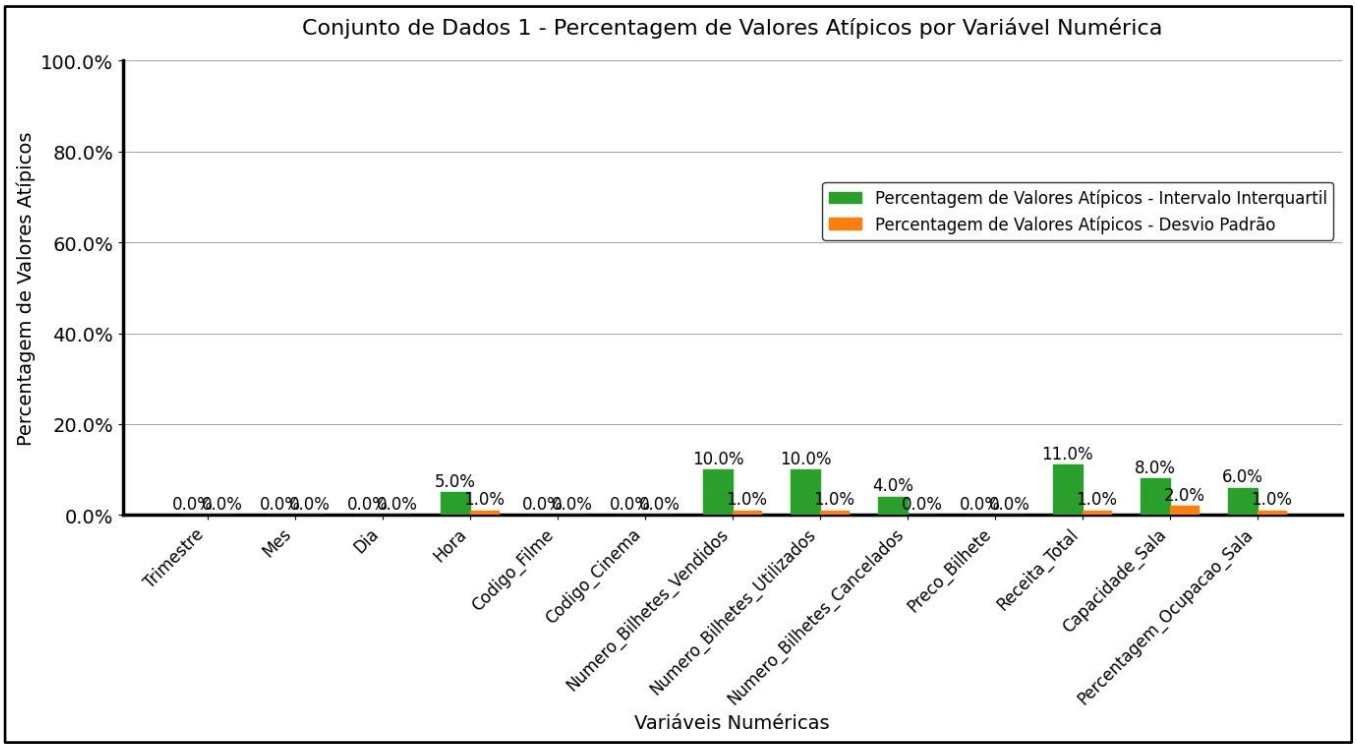

Figura d1.6 - Percentagem de Valores Atípicos por Variável Numérica.

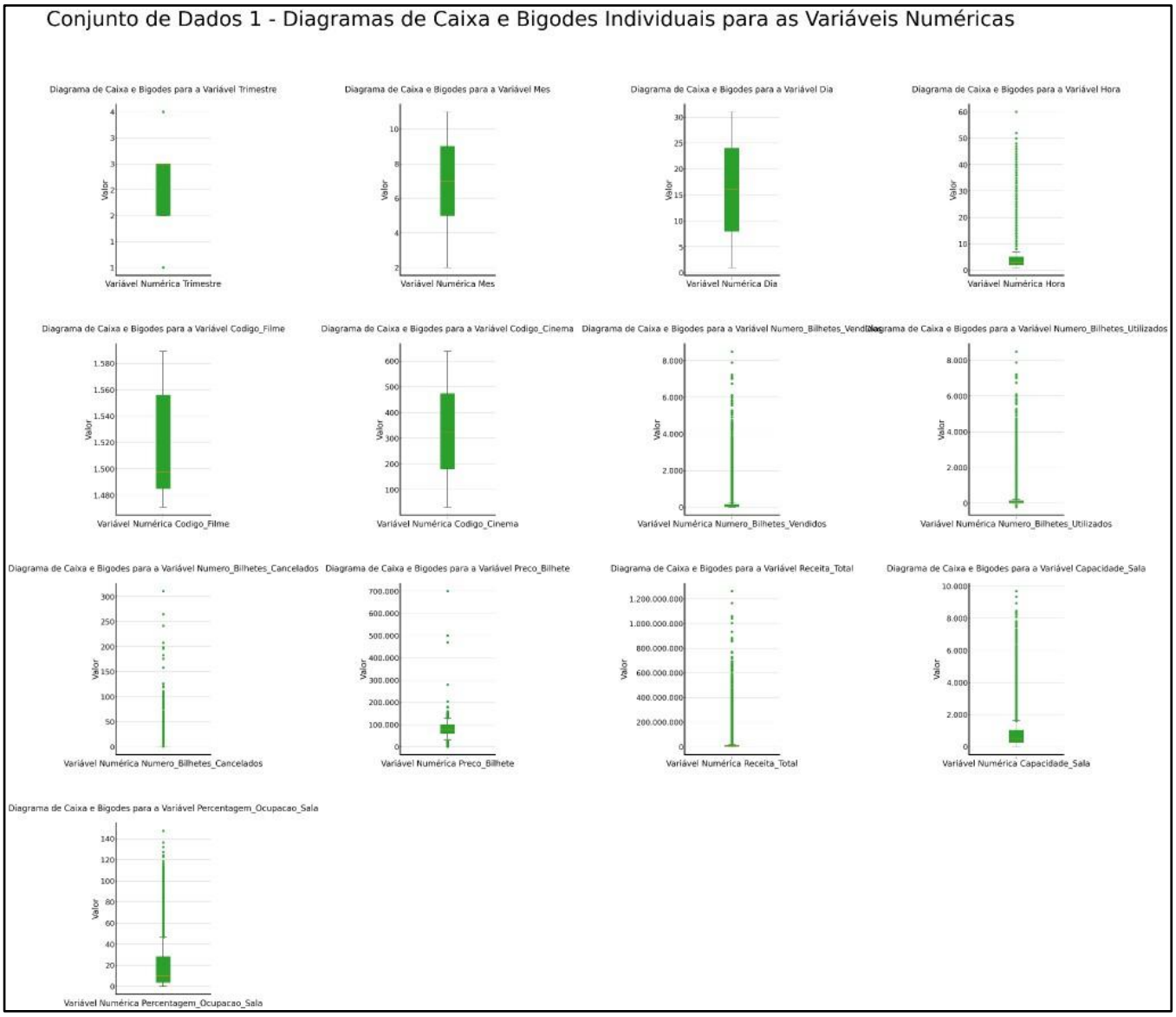

Figura d1.7 - Diagramas de Caixa e Bigodes Individuais para as Variáveis Numéricas.

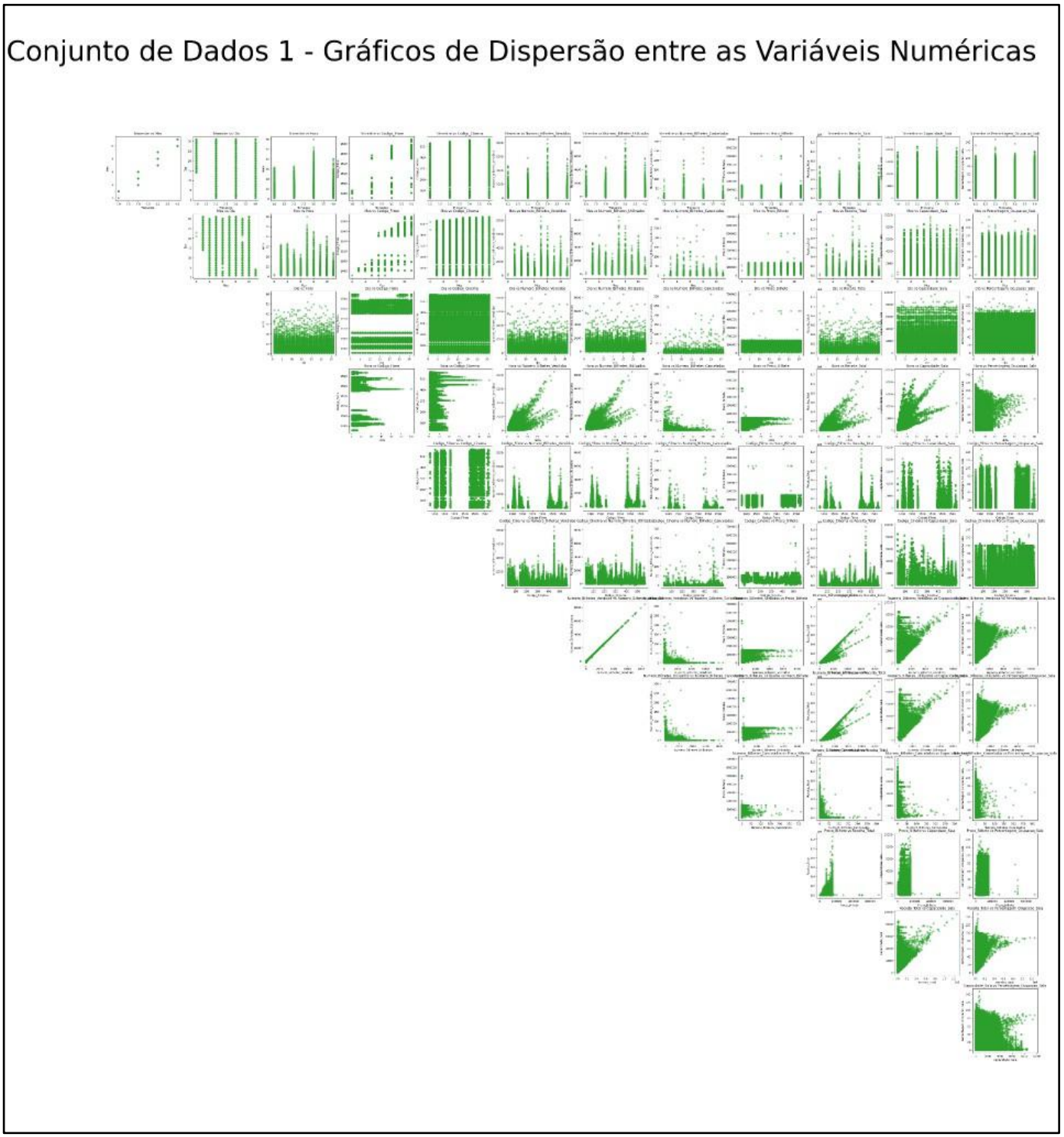

Figura d1.8 - Gráficos de Dispersão entre as Variáveis Numéricas.

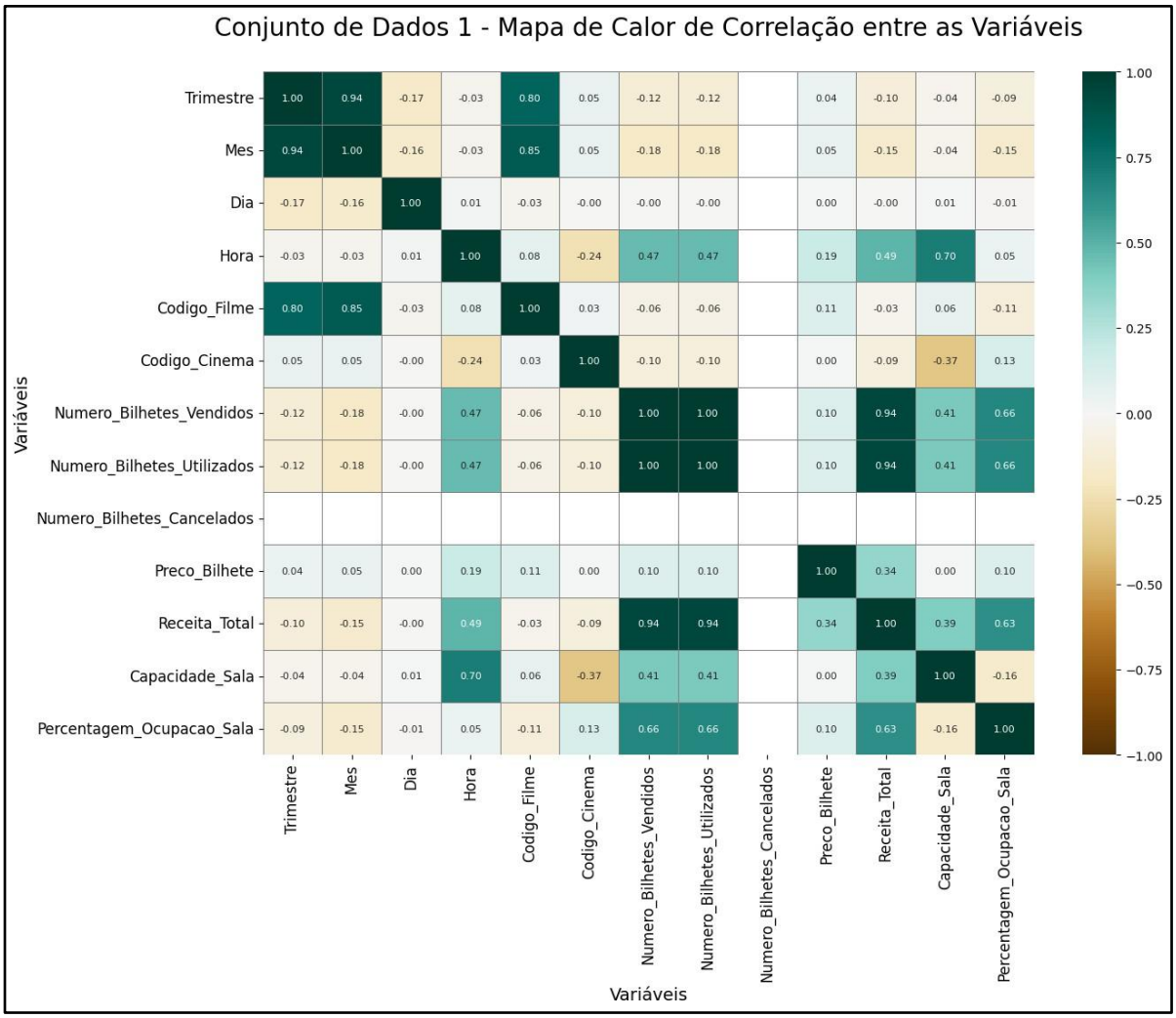

Figura d1.9 - Mapa de Calor de Correlação entre as Variáveis.

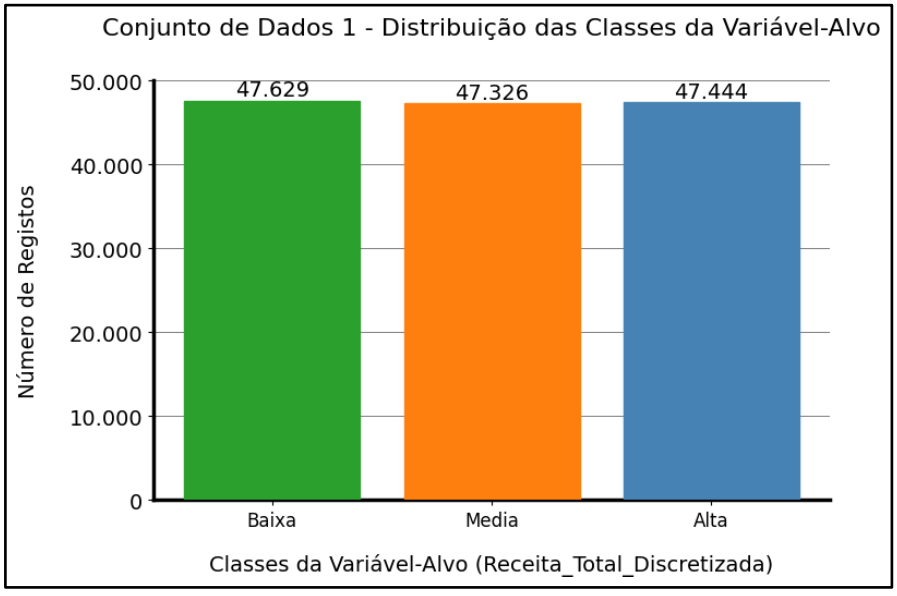

Figura d1.10 - Distribuição das Classes da Variável-Alvo (*Receita_Total_Discretizada*).

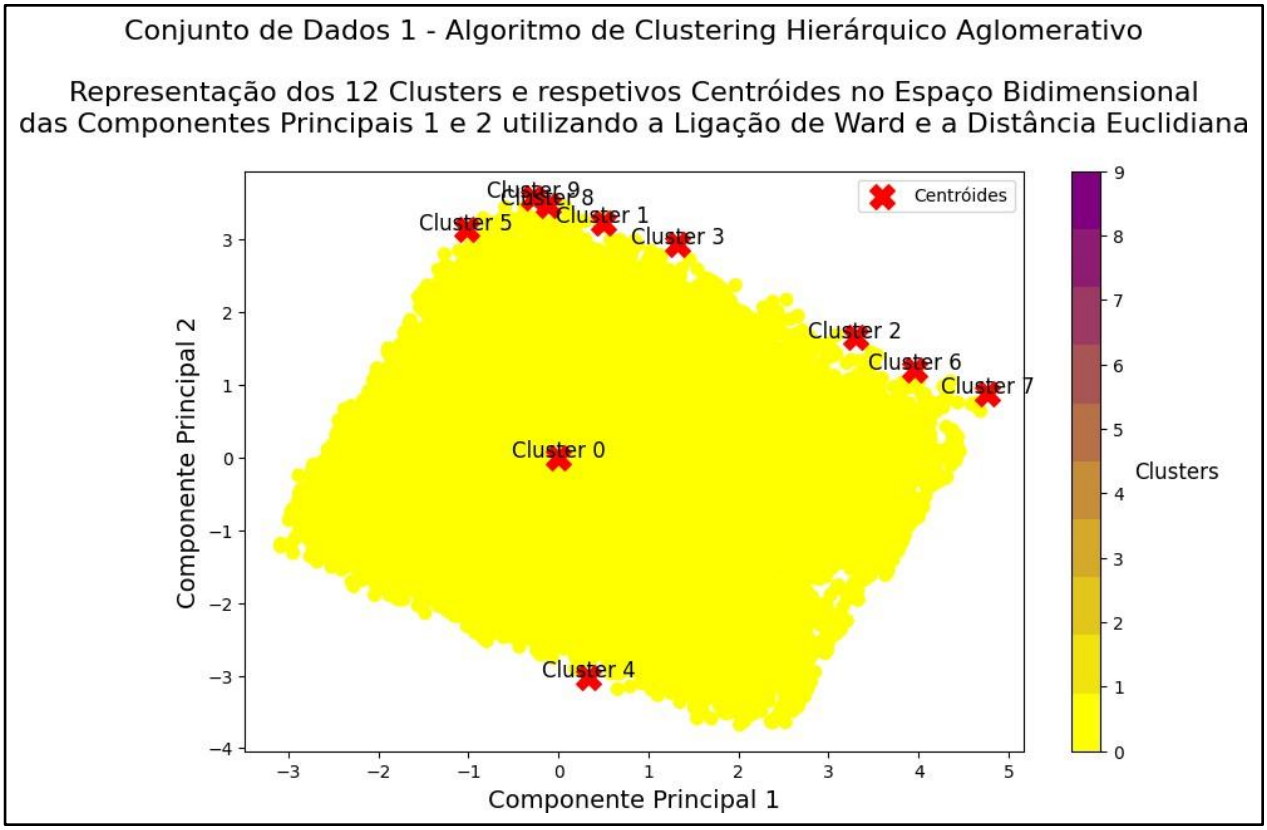

Figura d1.16 - Algoritmo de Clustering Hierárquico Aglomerativo - Representação dos 10 Clusters e respetivos Centróides nas Componentes Principais 1 e 2 utilizando a Ligação Mínima.