Relatório 2 *Aprendizado de Máquina*

Rafael Rosendo , Davi Almeida

Julho 2024

1 Introdução

O trabalho realizado na segunda unidade da disciplina de Aprendizado de Máquina tem como foco a aplicação de modelos de agrupamento, destacando-se os métodos de k-means, DBSCAN e a clusterização hierárquica. Todos os modelos foram aplicados a uma base de dados fornecida pelo professor, que aborda conceitos de segmentação de clientes, também conhecida como análise de cesta de compras. Esses algoritmos foram escolhidos por sua capacidade de agrupar objetos do dataset com base em características correspondentes, sendo métodos adequados para a divisão de grupos e separação de atributos.
Neste link se encontra o repositório com o código fonte

2 Preparação dos dados

Antes de aplicar os métodos de agrupamento, foi realizado um pré-processamento dos dados. Esse pré-processamento incluiu a remoção da coluna *CustomerID*, considerada desnecessária para os algoritmos de agrupamento, e a binarização dos valores de gênero, com Feminino: 0 e Masculino: 1. Em seguida, foi feita uma análise de correlação e uma visualização inicial dos dados do DataFrame, em que pode-se perceber a distribuição dos dados sobre as 3 dimensões destacadas, bem como o baixo impacto do gênero do cliente no posicionamento dos dados.

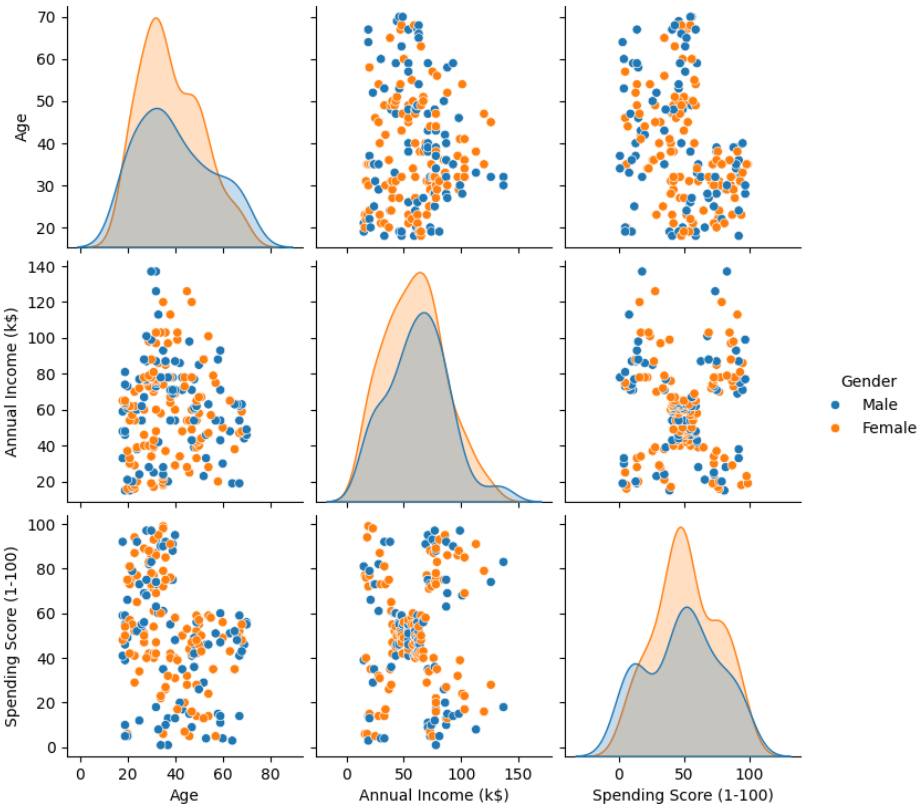

Figure 1: Distribuição dos dados do dataframe

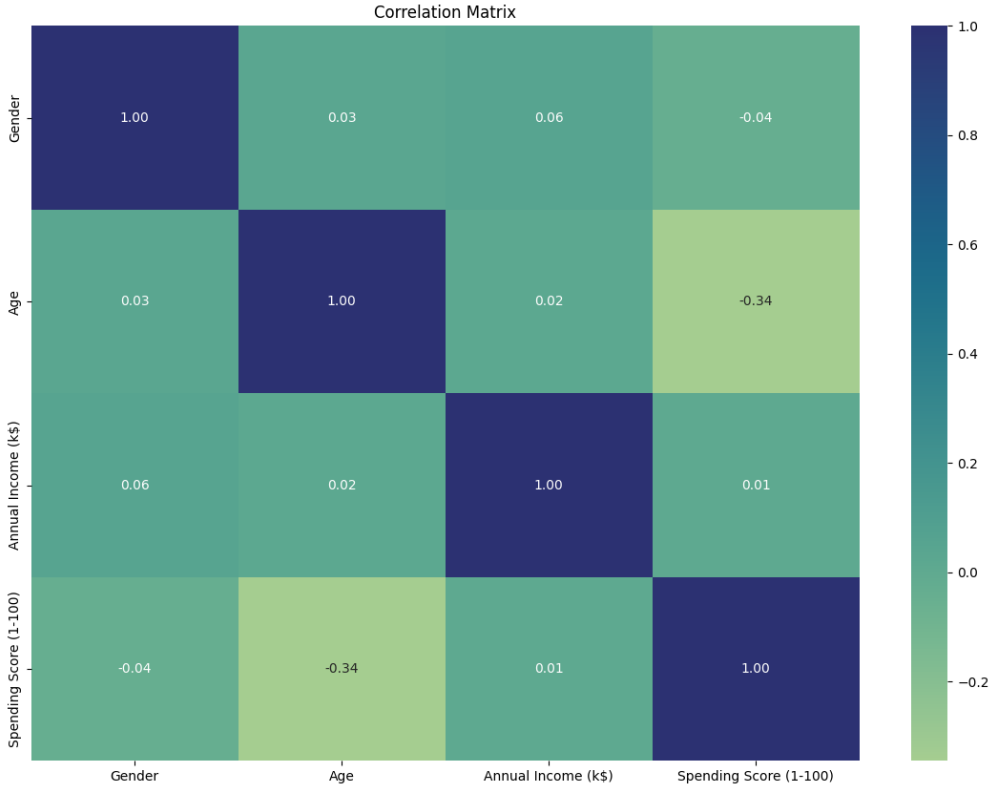

Figure 2: Correlação entre os dados do dataframe

3 Algoritmos de Agrupamento

Nesta seção, apresentaremos os três algoritmos de agrupamento utilizados neste estudo: K-Means, DBSCAN e Clustering Hierárquico. Cada um desses algoritmos tem suas próprias características e métodos de funcionamento, que são descritos a seguir.

3.1 K-Means

O algoritmo K-Means é um dos métodos de agrupamento mais populares e amplamente utilizados. Ele particiona os dados em k grupos, onde k é um número pré-definido de clusters. O algoritmo funciona iterativamente para atribuir cada ponto de dados ao cluster mais próximo, com base na média dos pontos no cluster.

Pontos Positivos:

- **Simplicidade e Eficiência**: É fácil de implementar e computacionalmente eficiente, especialmente para grandes datasets.
- **Escalabilidade**: Funciona bem com grandes datasets, com complexidade computacional linear em relação ao número de pontos.
- **Convergência Rápida**: Geralmente, converge rapidamente para um conjunto de clusters.
- **Clareza nos Resultados**: Os clusters são definidos de maneira clara e intuitiva, com cada ponto sendo atribuído ao cluster com o centróide mais próximo.

Pontos Negativos:

- **Número de Clusters Pré-Definido**: Requer que o número de clusters (k) seja definido antecipadamente.
- **Sensível a Outliers**: Outliers podem distorcer os centróides e impactar negativamente o resultado.
- **Forma dos Clusters**: Assume que os clusters são esféricos e de tamanho similar, o que nem sempre é o caso em dados reais.
- **Convergência a Mínimos Locais**: Pode convergir para mínimos locais, dependendo da inicialização dos centróides.

3.2 DBSCAN

O algoritmo DBSCAN (Density-Based Spatial Clustering of Applications with Noise) é um método de agrupamento baseado em densidade. Ele identifica clusters como regiões densamente povoadas de pontos de dados, separadas por regiões de menor densidade. Uma vantagem significativa do DBSCAN é sua capacidade de identificar pontos de dados como ruído (ou pontos sem grupo), se eles não pertencerem a nenhum cluster denso. Possui dois parâmetros que devem ser definidos: o 'eps'(epsilon) define a maior distância possível para que dois pontos sejam considerados vizinhos, ou seja, que possam fazer parte do mesmo cluster; o 'min_samples' define o número mínimo de amostras que devem ser agrupadas para a formação de um cluster.

Pontos Positivos:

- **Identificação de Outliers**: Identifica e separa pontos de ruído (outliers) de maneira eficaz.
- **Forma Arbitrária dos Clusters**: Pode identificar clusters de formas arbitrárias e de densidade variável.
- **Número de Clusters**: Não requer que o número de clusters seja definido antecipadamente.
- **Resistência a Outliers**: Menos influenciado por outliers, comparado ao K-Means.

Pontos Negativos:

- **Parâmetros Sensíveis**: A escolha dos parâmetros *eps* (raio de vizinhança) e *min_samples* (número mínimo de pontos) é crucial e pode ser difícil de ajustar corretamente.
- **Escalabilidade**: Pode ser computacionalmente intensivo para grandes datasets, especialmente se o *eps* for grande.
- **Clusters de Densidade Variável**: Pode ter dificuldades em identificar clusters se houver variação significativa na densidade dos dados.

3.3 Clustering Hierárquico

O Clustering Hierárquico é um método de agrupamento que constrói uma hierarquia de clusters e amostras. Existem duas abordagens principais: aglomerativa (de baixo para cima) e divisiva (de cima para baixo). Neste estudo, utilizamos a abordagem aglomerativa, onde cada ponto de dados começa em seu próprio cluster, e os clusters são sucessivamente combinados com base na proximidade até formar uma árvore hierárquica.

- **Não Requer Número de Clusters Pré-Definido**: Não precisa do número de clusters antecipadamente; os clusters podem ser determinados cortando o dendrograma em diferentes níveis.
- **Visualização Intuitiva**: O dendrograma oferece uma visualização clara da estrutura dos dados e das relações entre clusters.
- **Flexibilidade**: Pode utilizar diferentes critérios de ligação (single, complete, average, etc.), o que permite adaptar-se a diferentes formas de dados.
- **Identificação de Sub-Clusters**: Útil para identificar sub-estruturas hierárquicas nos dados.

Pontos Negativos:

- **Escalabilidade Limitada**: Computacionalmente intensivo e não escalável para grandes datasets, especialmente com critérios de ligação como *complete* e *average*.
- **Sensível a Ruído e Outliers**: Pode ser influenciado negativamente por ruídos e outliers, especialmente nos métodos aglomerativos.
- **Dificuldade em Atualizar**: Uma vez construído, o dendrograma não pode ser facilmente ajustado ou atualizado com novos dados.
- **Complexidade**: A complexidade computacional pode ser proibitiva para grandes conjuntos de dados.

3.3.1 Resumo Comparativo

- **K-Means**: Melhor para grandes datasets com clusters aproximadamente esféricos e de tamanho similar. Requer a definição prévia do número de clusters e é sensível a outliers.
- **DBSCAN**: Ideal para identificar clusters de forma arbitrária e densidade variável, além de lidar bem com outliers. Depende fortemente da escolha dos parâmetros e pode ser computacionalmente intensivo.
- **Clustering Hierárquico**: Útil para visualizar a estrutura dos dados e identificar sub-clusters. Não escala bem para grandes datasets e pode ser sensível a ruído e outliers.

4 Aplicação dos algoritmos e Resultados

Após o pré-processamento dos dados, aplicamos e analisamos diferentes algoritmos de agrupamento para identificar o melhor desempenho. Nesta seção, detalhamos os algoritmos utilizados, que incluem **K-Means**, **Clustering Hierárquico** e **DBSCAN**. Apresentamos seus respectivos resultados e incluímos gráficos e matrizes de confusão para facilitar a compreensão dos resultados obtidos.

4.1 K-Means

Inicialmente, foram realizados testes com valores de k entre 1 e 15, avaliando o erro entre grupos. O erro quadrático médio foi calculado e os gráficos gerados com base nesse erro e no agrupamento resultante do K-Means foram exibidos.

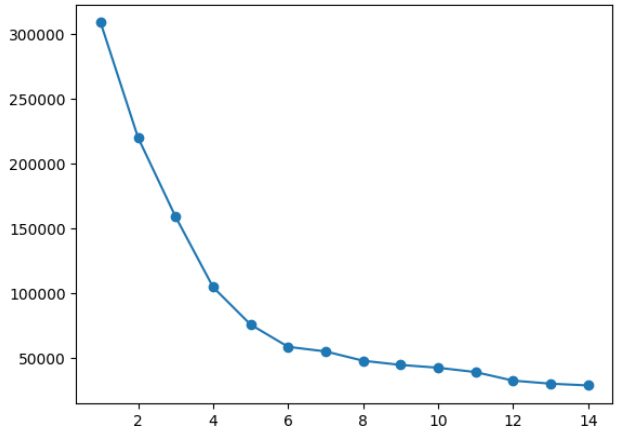

Figure 3: Erro quadrático médio ao longo de valores de k (1-14)

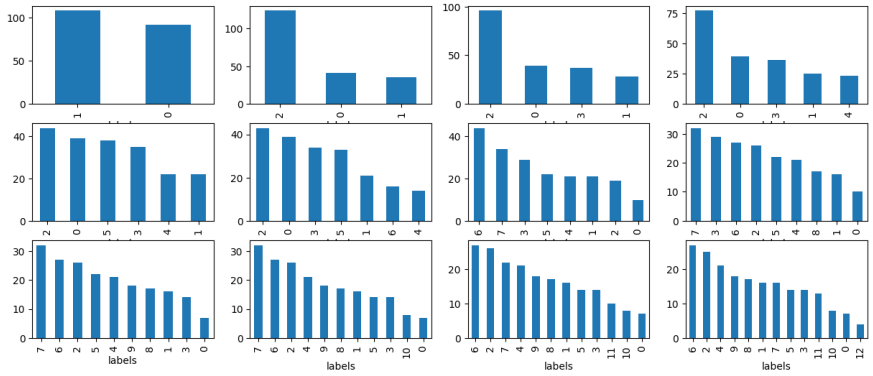

Figure 4: Número de clientes por cluster ao longo de valores de k (2-14)

Dados os resultados dos testes, foram feitas análises mais aprofundadas para dois valores de k específicos: 6 e 8. Para o valor k=8, cada cluster apresentou os seguintes valores médios:

Grupo	Gender	Age	Annual Income (k$)	Spending Score (1-100)
0	0.400000	32.200000	109.700000	82.000000
1	0.333333	27.047619	64.714286	44.952381
2	0.368421	27.631579	47.578947	53.842105
3	0.482759	32.862069	78.551724	82.172414
4	0.428571	25.333333	25.095238	80.047619
5	0.409091	44.318182	25.772727	20.272727
6	0.431818	56.340909	53.704545	49.386364
7	0.558824	41.647059	88.735294	16.764706

Table 1: Dados de agrupamento por grupo

Como é possível perceber, apesar da boa divisão em geral, alguns clusters apresentam muitas similaridades, como os clusters 1 e 2 e os clusters 0 e 3, além de índices Davies-Bouldin Index igual a 0.8646 e Silhouette Score igual a 0.396. Para k=6, obtivemos os seguintes valores médios para cada cluster:

Grupo	Gender	Age	Annual Income (k$)	Spending Score (1-100)
0	0.461538	32.692308	86.538462	82.128205
1	0.409091	25.272727	25.727273	79.363636
2	0.431818	56.340909	53.704545	49.386364
3	0.571429	41.685714	88.228571	17.285714
4	0.409091	44.318182	25.772727	20.272727
5	0.342105	27.000000	56.657895	49.131579

Table 2: Dados agregados por grupo

Com esse agrupamento, obtivemos índices Davies-Bouldin igual a 0.7449 e Silhouette Score igual a 0.452, ou seja, números consideravelmente melhores que na iteração anterior. A distribuição dos agrupamentos e a correlação entre as features são ilustradas abaixo:

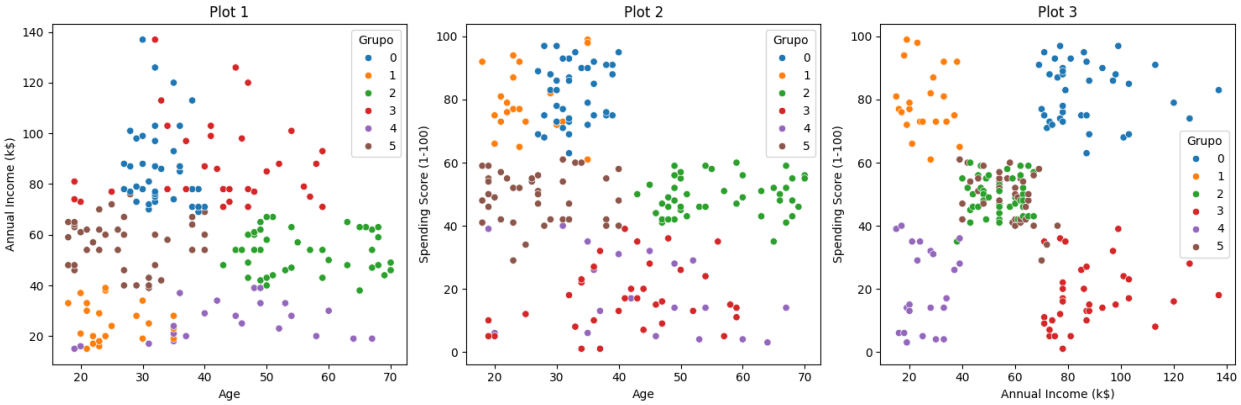

Figure 5: Distribuição dos grupos

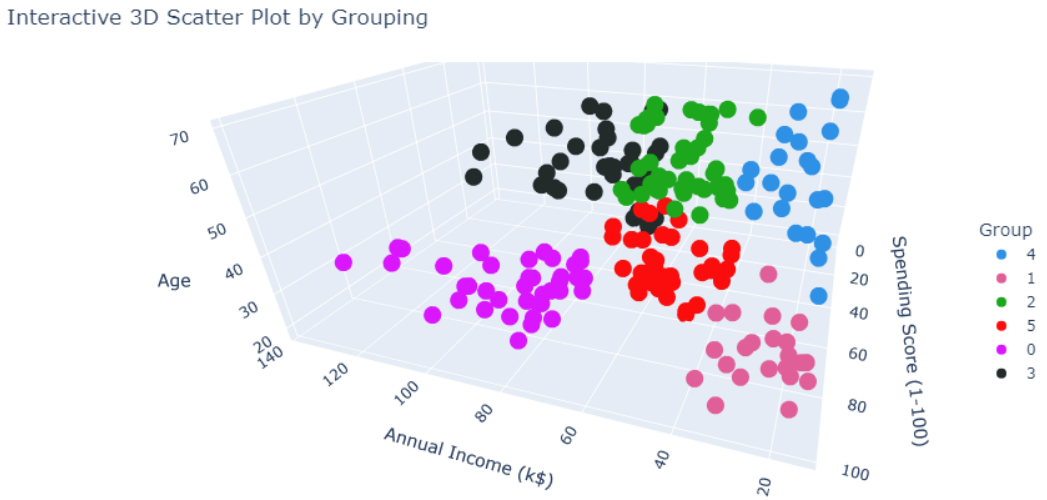

Figure 6: Distribuição dos grupos

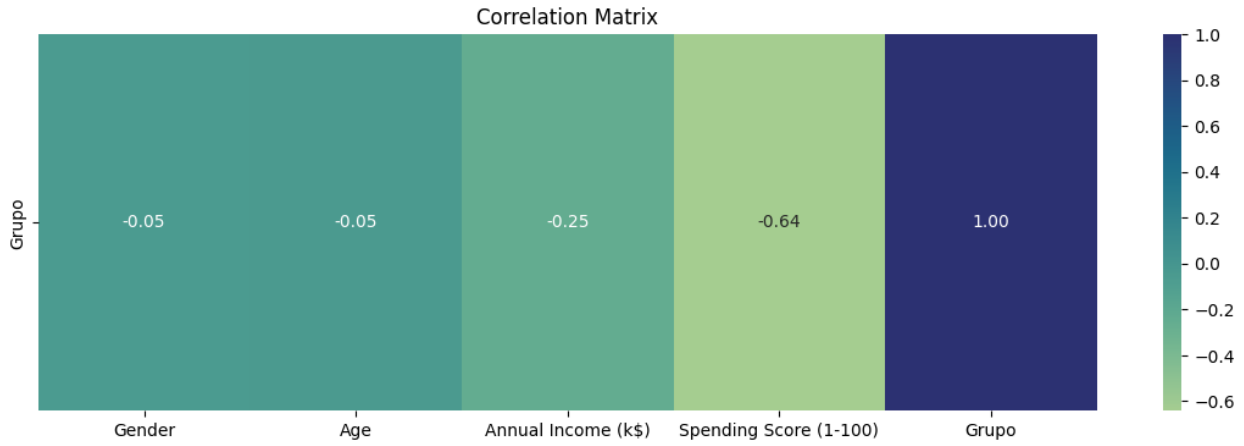

Figure 7: Correlação entre as métricas e o grupo atribuído

Os dados apresentados mostram que os fatores mais determinantes para a realização dos agrupamentos foram a renda anual e o Score do cliente, enquanto o gênero e a idade pouco afetaram o resultado final.

Realizando uma análise empírica dos 6 agrupamentos criados, temos os seguintes resultados:

Grupo 0

- Gênero predominante: Não há
- Idade: Média de 32.7 anos (adultos jovens)
- Renda Anual: $86.5k (renda alta)
- Pontuação de Gastos: 82.1 (alta pontuação de gastos)

Descrição: Este grupo possui uma renda elevada e uma alta pontuação de gastos, indicando que são clientes que investem significativamente na loja. A faixa etária é jovem, sugerindo que são profissionais em início de carreira ou jovens adultos com boa capacidade de gasto.

Grupo 1

- Gênero predominante: Não há
- Idade: Média de 25.3 anos (jovens adultos)
- Renda Anual: $25.7k (renda baixa)
- Pontuação de Gastos: 79.4 (alta pontuação de gastos)

Descrição: Apesar da renda ser relativamente baixa, este grupo ainda tem uma alta pontuação de gastos, o que pode indicar que são jovens adultos que gastam de forma relativamente generosa em comparação com sua renda. Eles podem priorizar o gasto em itens de alto valor ou estar dispostos a gastar mais do que sua renda sugere.

Grupo 2

- Gênero predominante: Não há
- Idade: Média de 56.3 anos (adultos mais velhos)
- Renda Anual: $53.7k (renda média)
- Pontuação de Gastos: 49.4 (pontuação de gastos moderada)

Descrição: Este grupo é composto por clientes mais velhos com uma renda média e gastos moderados. Eles podem estar em uma fase de vida onde são menos propensos a gastar grandes quantias, optando por um consumo mais controlado.

Grupo 3

- Gênero predominante: Não há
- Idade: Média de 41.7 anos (adultos de meia-idade)
- Renda Anual: $88.2k (renda alta)
- Pontuação de Gastos: 17.3 (baixa pontuação de gastos)

Descrição: Este grupo tem uma renda alta, mas uma baixa pontuação de gastos. Isso sugere que, embora tenham alta capacidade financeira, são clientes que gastam pouco na loja, possivelmente comprando menos ou preferindo itens mais baratos.

Grupo 4

- Gênero predominante: Não há
- Idade: Média de 44.3 anos (adultos de meia-idade)
- Renda Anual: $25.8k (renda baixa)
- Pontuação de Gastos: 20.3 (baixa pontuação de gastos)

Descrição: Este grupo possui uma baixa renda e baixa pontuação de gastos. Eles podem ser clientes que têm restrições financeiras e gastam pouco, o que pode indicar um padrão de consumo muito controlado.

Grupo 5

- Gênero predominante: Não há
- Idade: Média de 27 anos (jovens adultos)
- Renda Anual: $56.7k (renda média)
- Pontuação de Gastos: 49.1 (pontuação de gastos moderada)

Descrição: Este grupo é composto por jovens adultos com renda média e uma pontuação de gastos moderada. Eles estão em um momento onde têm uma capacidade financeira razoável e gastam de forma equilibrada.

4.2 Clustering Hierárquico

Foi utilizado um modelo de clustering hierárquico configurado para calcular a árvore completa (*distance-threshold=0*), sem um número pré-definido de clusters (*n-clusters=None*). Para a divisão dos grupos, foram utilizados 13 subgrupos.

Inicialmente, vamos apresentar algumas imagens para exemplificar melhor e abordar com mais propriedade os resultados obtidos a partir do algoritmo de cluster hierárquico. A primeira imagem mostrará o número de clientes por cluster para valores entre 2 e 13.

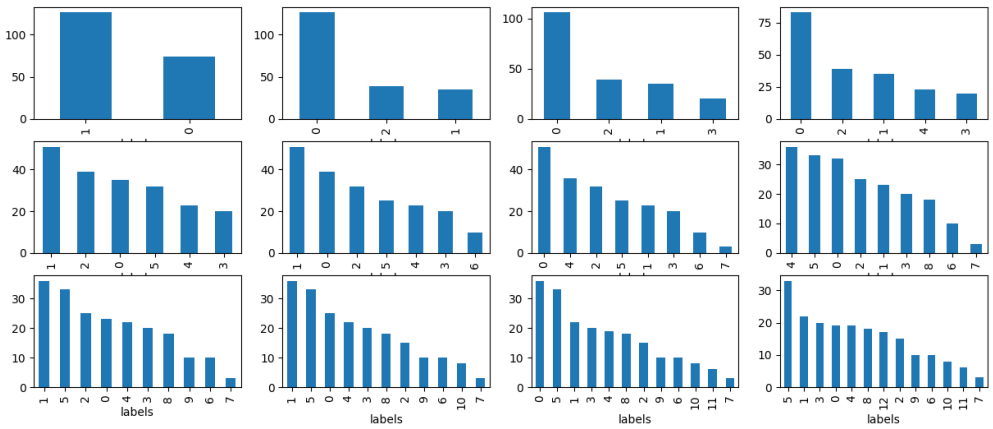

Figure 8: Distribuição dos grupos dado um K

Logo depois de fazer uma separação por agrupamentos, podemos então trazer a tona algumas imagens de algumas métricas usadas em atividades de clusterização

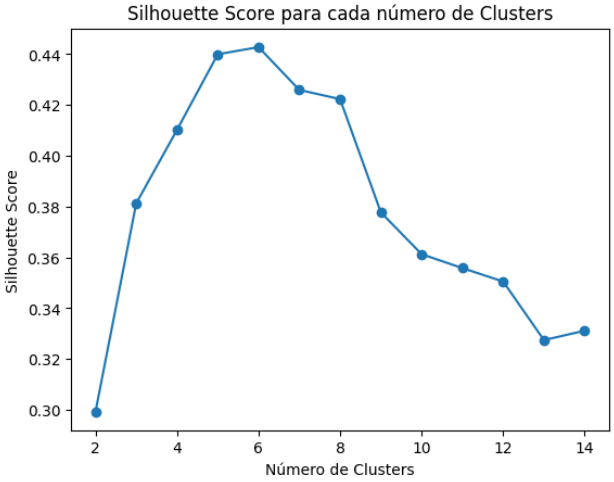

Figure 9: Curva de Silhouette

Após essa análise, é possível perceber que, utilizando o referido algoritmo, o número de clusters ideal será 6. Partindo para outro ponto de vista podemos também ver a clusterização com outros olhos e implementar uma imagem que demonstre a criação de um dendograma, como pode notar logo a baixo:

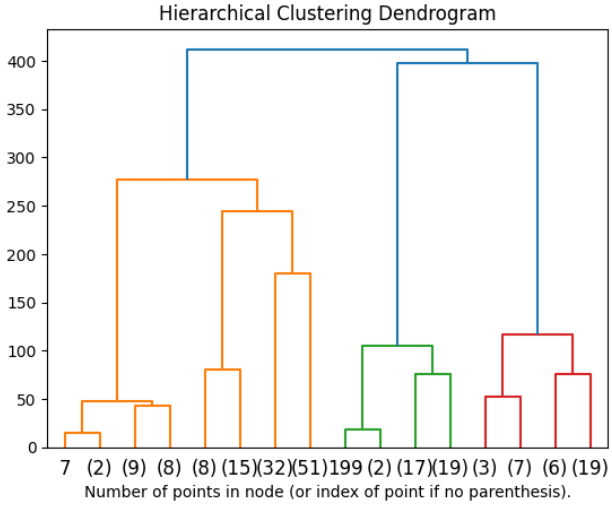

Figure 10: Dendograma

Para entender de outra forma o cenário, podemos criar uma tabela usando os dados gerados pelo modelo e assim ficamos com:

Grupo	Gender	Age	Annual Income (k$)	Spending Score (1-100)
0	0.571429	41.685714	88.228571	17.285714
1	0.411765	53.215686	55.117647	49.470588
2	0.461538	32.692308	86.538462	82.128205
3	0.400000	24.850000	24.950000	81.000000
4	0.391304	45.217391	26.304348	20.913043
5	0.375000	24.531250	54.187500	50.250000

Table 3: Dados agregados por grupo com 6 clusters

Agora que vimos como o modelo se comporta, podemos ter uma ideia mais gráfica e visual usando alguns gráficos como referência, onde podemos visualizar a divisão de grupos representados em uma dimensão (X, Y) para ver como está a distribuição de grupos

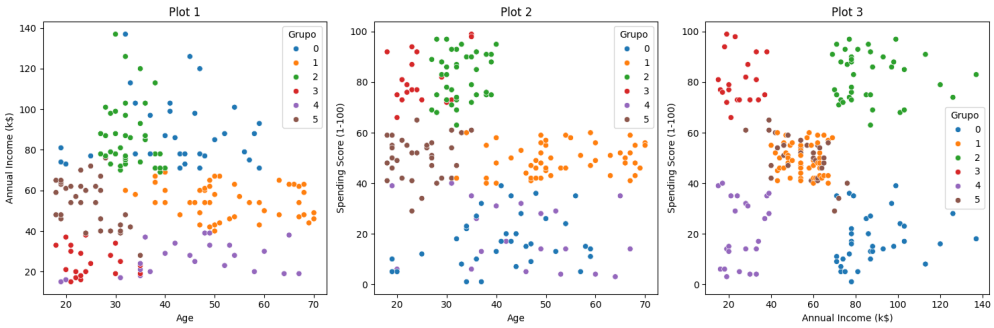

Figure 11: Distribuição dos Grupos

Dados os clusters apresentados, realizaremos uma análise empírica sobre a divisão alcançada com o algoritmo Cluster Hierárquico:

Grupo 0

- Gênero predominante: Não há
- Idade: Média de 41.6 anos (adultos)
- Renda Anual: $88.2k (renda alta)
- Pontuação de Gastos: 17 (baixa pontuação de gastos)

Descrição: Este grupo é composto por adultos com renda alta, e que possuem uma baixa pontuação de gastos. Isso pode indicar que, apesar de terem uma renda relativamente boa, vê o padrão de gastarem pouco

Grupo 1

- Gênero predominante: Não há
- Idade: Média de 53.2 anos (adultos de meia-idade)
- Renda Anual: $55.1k (renda média)
- Pontuação de Gastos: 48.8 (pontuação média de gastos)

Descrição: Este grupo é formado por adultos de meia-idade com uma renda média e uma pontuação de gastos média. Eles podem ser consumidores estáveis, com hábitos de gasto conservadores, possivelmente priorizando economia.

Grupo 2

- Gênero predominante: Não há
- Idade: Média de 32 anos (adultos)
- Renda Anual: $86.5k (renda baixa)
- Pontuação de Gastos: 82 (pontuação alta de gastos)

Descrição: Este grupo é composto por adultos com uma renda alta e uma pontuação de gastos alta. Eles parecem ser bons consumidores, com um comportamento de gasto mais condizente em relação à sua renda.

Grupo 3

- Gênero predominante: Não há
- Idade: Média de 24.8 anos (jovens adultos)

- Renda Anual: $24.9k (renda baixa)
- Pontuação de Gastos: 50.0 (pontuação média de gastos)

Descrição: Este grupo é composto principalmente por jovnes com uma renda baixa e uma pontuação de gastos alta em relação ao salário dos mesmos. Eles parecem ser consumidores ferozes, possivelmente focando em muitas compras.

Grupo 4

- Gênero predominante: Maioria de mulheres (65%)
- Idade: Média de 45.2 anos (adultos)
- Renda Anual: $26.3k (renda baixa)
- Pontuação de Gastos: 20.0 (alta pontuação de gastos)

Descrição: Este grupo é composto por adultos com uma renda baixa e uma baixa pontuação de gastos. Eles são consumidores pacientes, provavelmente comprando somente o necessário.

Grupo 5

- Gênero predominante: Somente Masculino
- Idade: Média de 24.5 anos (jovens adultos)
- Renda Anual: $54k (renda média)
- Pontuação de Gastos: 50 pontuação de gastos moderada)

Descrição: Este grupo é composto exclusivamente por homens jovens com uma renda média-alta, e com uma pontuação condizente com seu ganho. Isso pode indicar que, apesar da capacidade financeira, elas têm um comportamento de consumo moderado.

Grupo 6

- Gênero predominante: Não há
- Idade: Média de 43.2 anos (adultos de meia-idade)
- Renda Anual: $81.4k (renda alta)
- Pontuação de Gastos: 17.2 (baixa pontuação de gastos)

Descrição: Este grupo é composto por adultos de meia-idade com uma renda alta, mas uma baixa pontuação de gastos. Eles parecem ser consumidores muito cuidadosos e conservadores.

4.3 DBSCAN

Em seguida, foi implementado o DBSCAN. De início, foram feitas análises acerca dos índices Silhouette Score e Davies-Bouldin Index sobre uma variedade de valores para os parâmetros eps e min_samples.

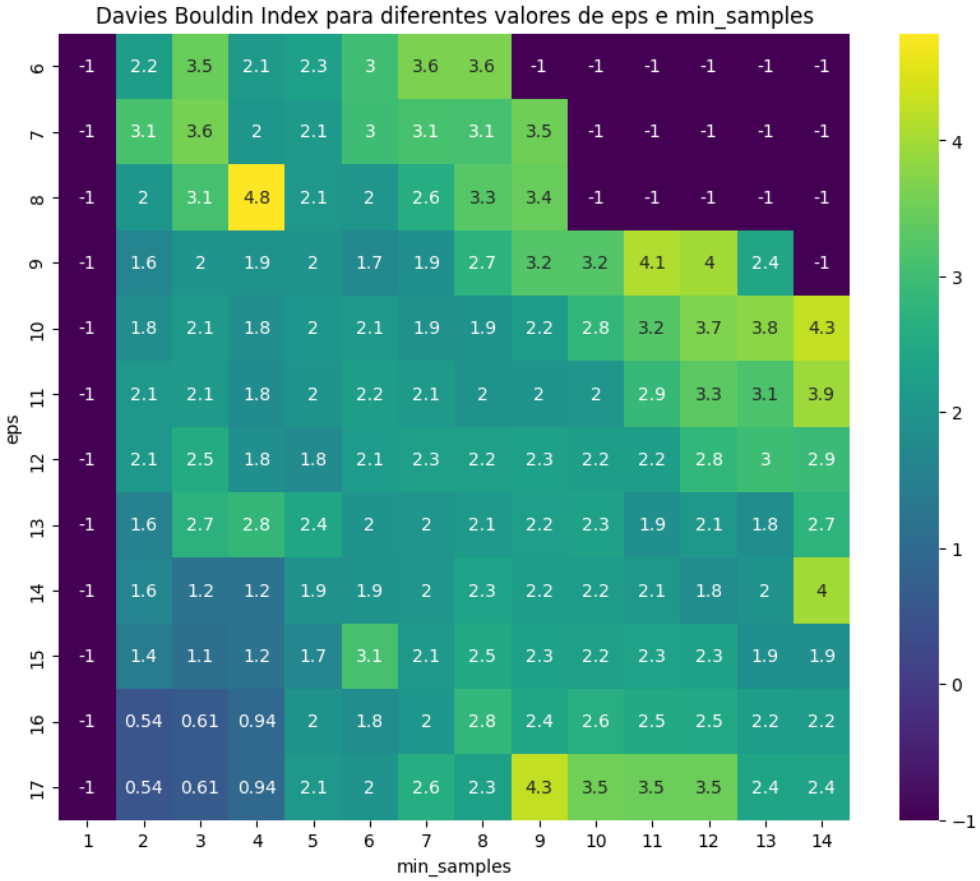

Figure 12: Heatmap do Davies-Bouldin Index

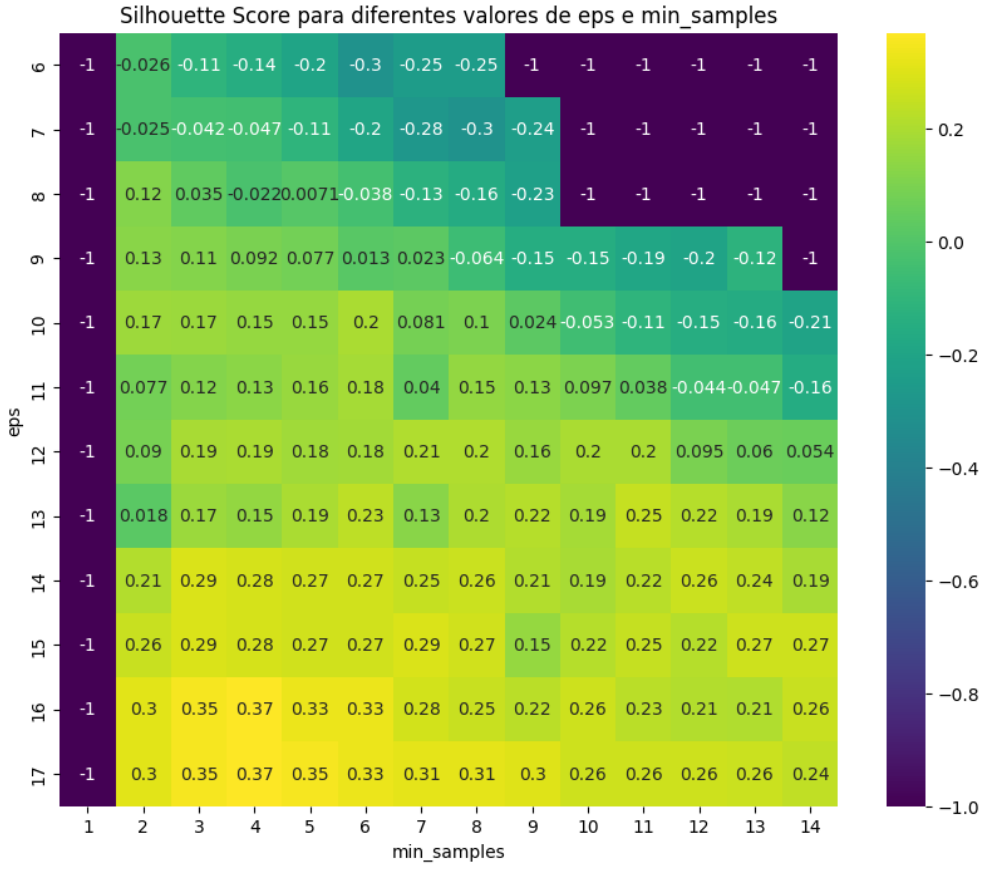

Figure 13: Heatmap do Silhouette Score

Realizando uma análise que leva em consideração unicamente as métricas apresentadas, fica claro, pelo gradiente, que elas apresentam uma melhoria com valores mais baixos de min_samples e diretamente proporcional ao valor de eps. Porém, precisamos analisar se essa melhoria de índice apresenta uma melhoria real na criação de clusters que sejam úteis para o vendedor. Para isso, definimos o min_samples para 4 e iteramos sobre valores de eps de 6 a 17, obtendo a seguinte divisão de grupos:

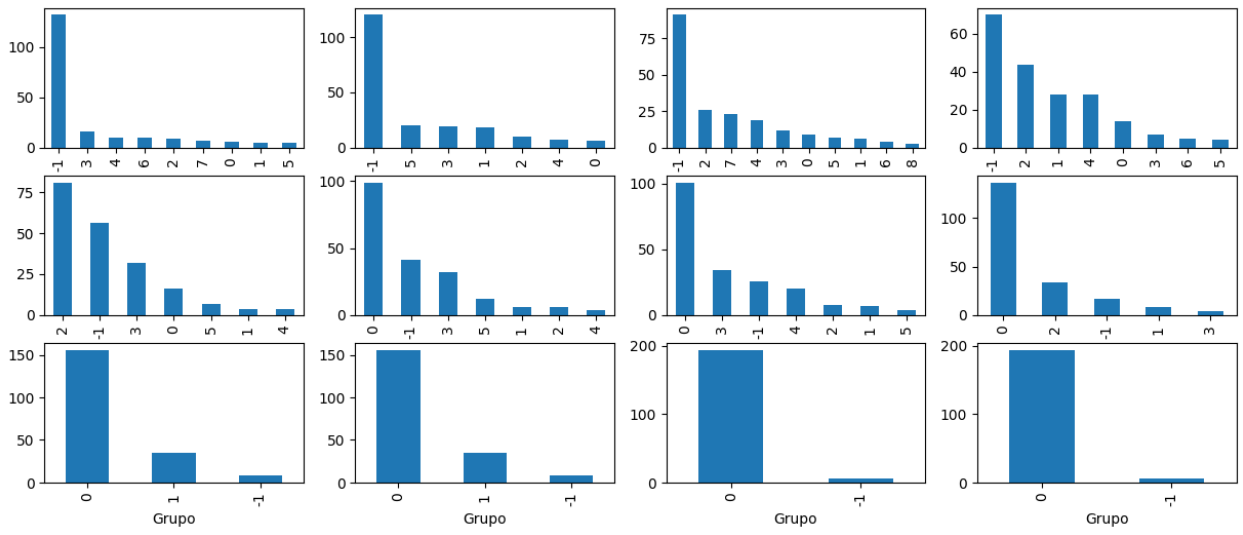

Figure 14: Distribuição dos Grupos dado min_samples=4 e valor eps variável

Com isso, é possível perceber que a busca por índices ideais tem como consequência uma divisão com apenas um ou dois grupos, o que não é interessante para nosso caso de estudo. Vamos visualizar a clusterização computada pelo algoritmo com uma série de parâmetros diferentes:

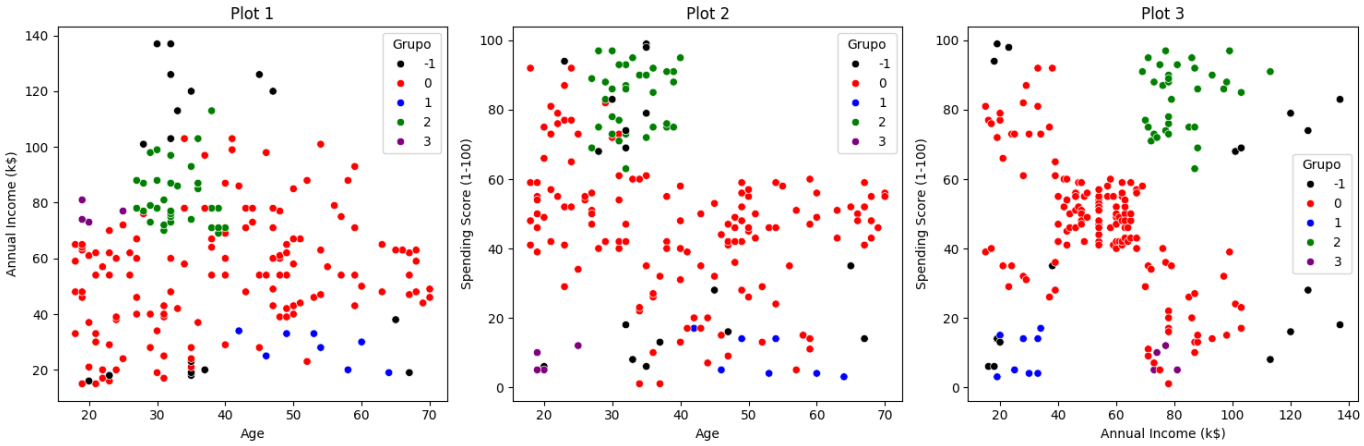

Figure 15: eps=10, min_samples=10

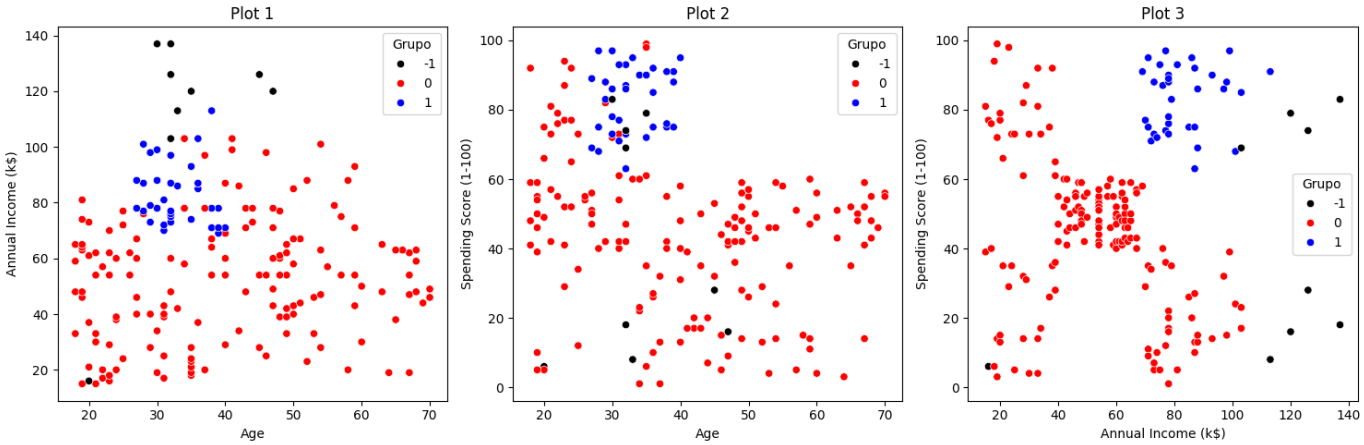

Figure 16: eps=14, min_samples=4

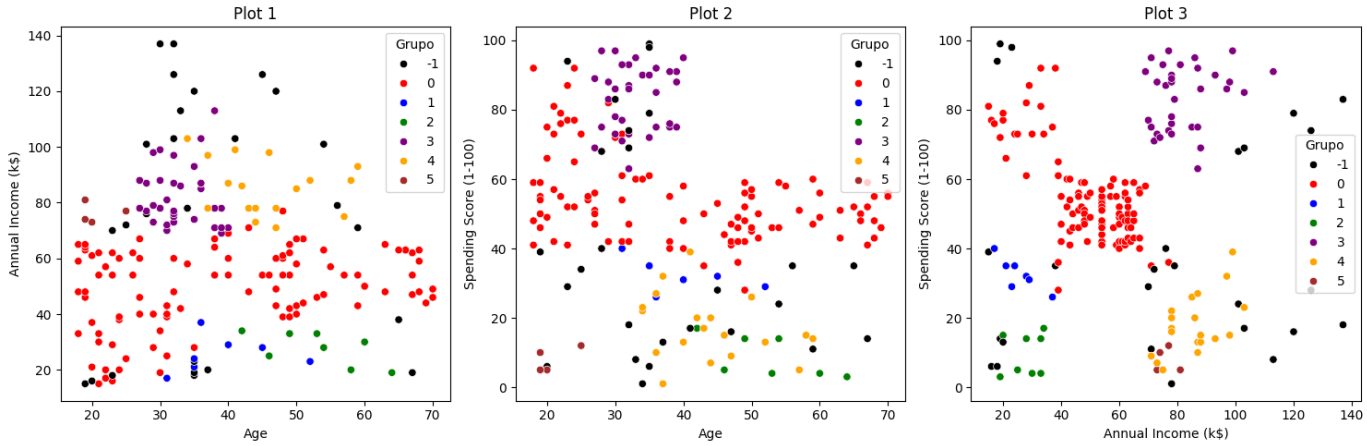

Figure 17: eps=12, min_samples=4

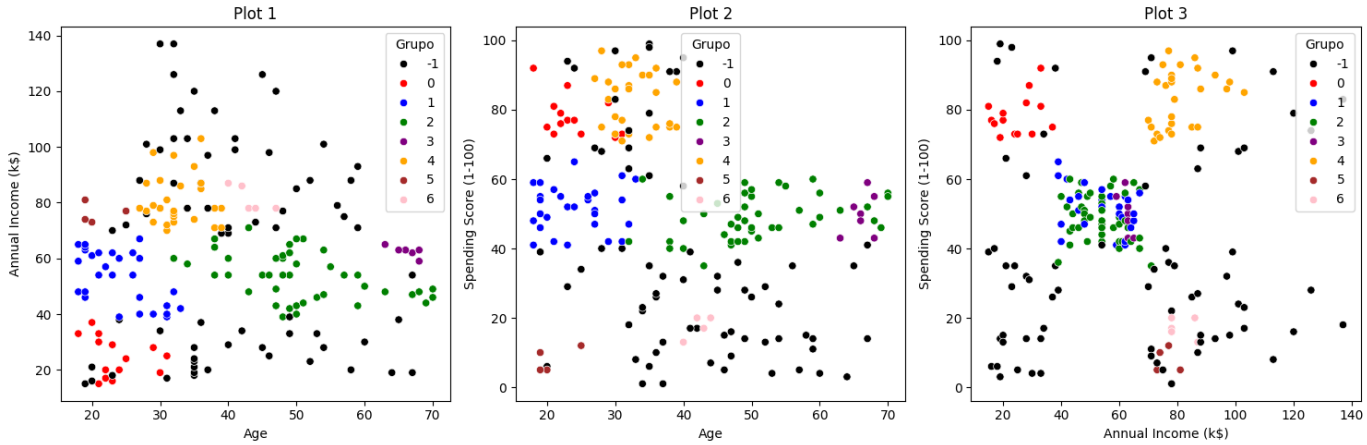

Figure 18: eps=9, min_samples=4

Temos que a diminuição do eps aumenta a quantidade de amostras identificadas como outlier, enquanto o aumento desse parâmetro gera grupos maiores, com menos outliers. No geral, o algoritmo não apresentou clusters e índices muito satisfatórios, porém vamos aprofundar a análise no agrupamento realizado com parâmetros eps=9 e min_samples=4, que, apesar de possuir 70 amostras marcadas como outliers, apresenta 7 grupos de clientes bem definidos e separados, o que pode ser relevante considerando nosso objetivo final.

Vamos, inicialmente, observar a média dos features de cada cluster, bem como uma visualização em 3d e a correlação de pearson entre os grupos e as features:

Grupo	Gender	Age	Annual Income (k$)	Spending Score (1-100)
-1	0.457143	40.814286	64.385714	36.585714
0	0.428571	23.571429	24.714286	78.428571
1	0.392857	24.071429	53.142857	51.571429
2	0.409091	50.795455	53.636364	48.750000
3	0.571429	66.142857	62.571429	50.000000
4	0.357143	32.571429	80.750000	83.000000
5	1.000000	20.750000	76.250000	8.000000
6	0.600000	43.200000	81.400000	17.200000

Table 4: Valor médio de cada feature nos clusters gerados. eps=9, min_samples=4

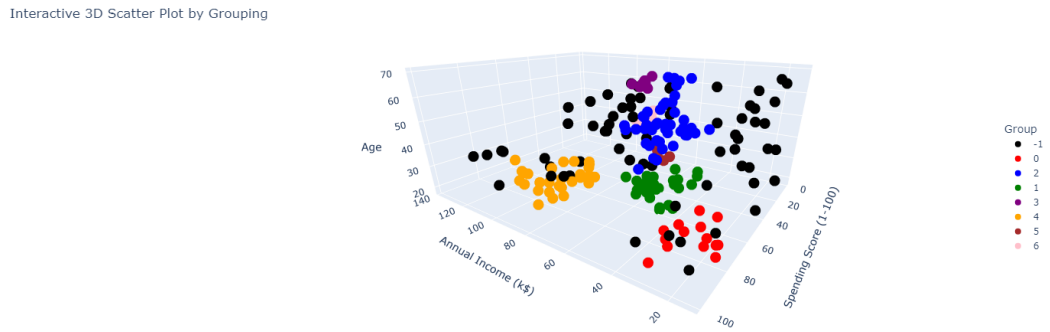

Figure 19: Scatterplot 3d. eps=9, min_samples=4

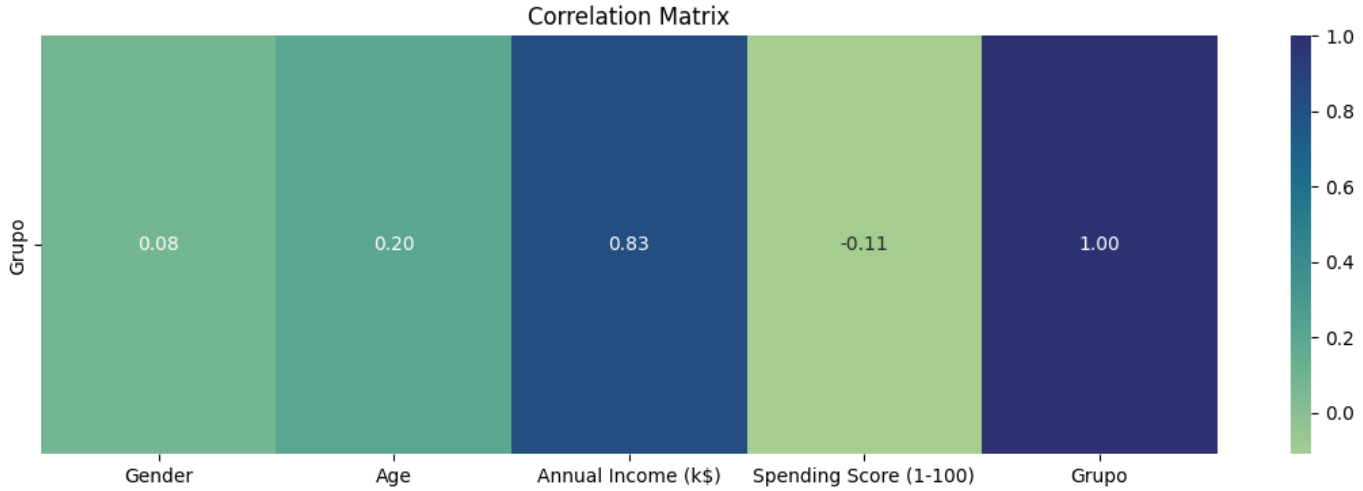

Figure 20: Correlação desconsiderando os outliers. eps=9, min_samples=4

Dados os clusters apresentados, realizaremos uma análise empírica sobre a divisão alcançada com o algoritmo DBSCAN:

Grupo 0

- Gênero predominante: Não há
- Idade: Média de 23.6 anos (jovens adultos)
- Renda Anual: $24.7k (renda baixa)
- Pontuação de Gastos: 78.4 (alta pontuação de gastos)

Descrição: Este grupo é composto por jovens adultos com renda baixa, mas que têm uma alta pontuação de gastos. Isso pode indicar que, apesar de terem uma renda limitada, eles tendem a gastar uma boa parte dela no estabelecimento.

Grupo 1

- Gênero predominante: Não há
- Idade: Média de 24.1 anos (jovens adultos)
- Renda Anual: $53.1k (renda média)

- Pontuação de Gastos: 51.6 (pontuação média de gastos)

Descrição: Este grupo é composto por jovens adultos com uma renda média e uma pontuação de gastos também média. Eles parecem ser consumidores moderados, com um comportamento de gasto equilibrado em relação à sua renda.

Grupo 2

- Gênero predominante: Não há
- Idade: Média de 50.8 anos (adultos de meia-idade)
- Renda Anual: $53.6k (renda média)
- Pontuação de Gastos: 48.8 (pontuação média de gastos)

Descrição: Este grupo é formado por adultos de meia-idade com uma renda média e uma pontuação de gastos média. Eles podem ser consumidores estáveis, com hábitos de gasto conservadores, possivelmente priorizando economia.

Grupo 3

- Gênero predominante: Não há
- Idade: Média de 66.1 anos (idosos)
- Renda Anual: $62.6k (renda média-alta)
- Pontuação de Gastos: 50.0 (pontuação média de gastos)

Descrição: Este grupo é composto principalmente por idosos com uma renda média-alta e uma pontuação de gastos média. Eles parecem ser consumidores moderados, possivelmente focando em compras de necessidade.

Grupo 4

- Gênero predominante: Maioria de mulheres (65%)
- Idade: Média de 32.6 anos (adultos jovens)
- Renda Anual: $80.8k (renda alta)
- Pontuação de Gastos: 83.0 (alta pontuação de gastos)

Descrição: Este grupo é composto por adultos jovens com uma renda alta e uma alta pontuação de gastos. Eles são consumidores significativos, provavelmente investindo em produtos de alta qualidade ou luxo.

Grupo 5

- Gênero predominante: Somente Masculino
- Idade: Média de 20.8 anos (jovens adultos)
- Renda Anual: $76.3k (renda alta)
- Pontuação de Gastos: 8.0 (baixa pontuação de gastos)

Descrição: Este grupo é composto exclusivamente por homens jovens com uma renda alta, mas uma baixa pontuação de gastos. Isso pode indicar que, apesar da capacidade financeira, elas têm um comportamento de consumo muito conservador.

Grupo 6

- Gênero predominante: Não há
- Idade: Média de 43.2 anos (adultos de meia-idade)
- Renda Anual: $81.4k (renda alta)
- Pontuação de Gastos: 17.2 (baixa pontuação de gastos)

Descrição: Este grupo é composto por adultos de meia-idade com uma renda alta, mas uma baixa pontuação de gastos. Eles parecem ser consumidores muito cuidadosos e conservadores.

5 Conclusão e possíveis aplicações

Neste trabalho, exploramos e comparamos três métodos populares de clustering: K-Means, DBSCAN e Clustering Hierárquico. Cada método possui suas vantagens e desvantagens, o que os torna adequados para diferentes tipos de problemas e conjuntos de dados.

O K-Means é eficiente e escalável, ideal para grandes datasets com clusters aproximadamente esféricos. No entanto, requer a definição prévia do número de clusters e é sensível a outliers. O DBSCAN, por outro lado, é robusto contra outliers e pode identificar clusters de forma arbitrária e densidade variável, mas depende fortemente da escolha de parâmetros e pode ser computacionalmente intensivo para grandes datasets. Já o Clustering Hierárquico oferece uma visualização clara da estrutura dos dados e não necessita da definição prévia do número de clusters, mas é menos escalável e sensível a ruídos e outliers.

As possíveis aplicações desses métodos são vastas. O K-Means é amplamente utilizado em segmentação de clientes, compressão de imagens e análise de padrões. O DBSCAN é útil em detecção de anomalias, reconhecimento de padrões espaciais e análise de grandes conjuntos de dados com ruído. O Clustering Hierárquico é aplicado em bioinformática para construir árvores filogenéticas, análise de textos e mineração de dados para descobrir subestruturas hierárquicas.

Na Aplicação específica estudada, o K-Means e o Clustering Hierárquico apresentaram resultados mais satisfatórios, enquanto o DBSCAN, devido à ausência de grupamentos de alta densidade no dataset devido à baixa quantidade de dados, precisou de um maior refino para alcançar uma estratégia de agrupamento considerado suficiente.

Em conclusão, a escolha do algoritmo de clustering deve ser guiada pelas características específicas do dataset e pelos objetivos da análise, garantindo a seleção da abordagem mais adequada para obter insights significativos.