**UNIVERSIDADE FEDERAL DE SÃO CARLOS**
Centro de Ciências Exatas e Tecnologia
Departamento de Computação

# TRABALHO PRÁTICO 2 - AGRUPAMENTO NÃO-SUPERVISIONADO

**Aprendizado de Máquina 1**
Prof.: Murilo Coelho Naldi

Jayme Sakae dos Reis Furuyama - 761044
Vitor Lopes Fabris - 769822

16 de Setembro de 2022

# Índice

# 1. Introdução

Nos dias atuais, a quantidade de dados tem se multiplicado de forma avassaladora. A navegação na web, a realização de transações bancárias, a administração de empresas, etc., geram cotidianamente centenas de milhares de informações novas.

O aprendizado de máquina se debruça sobre o problema de dar significado a esses dados. Muitas vezes, dados especiais (chamados de dados de treinamento) são separados para diferentes tarefas com rótulos, ou seja, significados embutidos nas informações antecipadamente. Porém, com essa geração constante de dados novos e o custo elevado de rotular conjuntos grandes de dados, é necessário aplicar técnicas que introduzem significado a conjuntos aparentemente sem nenhum. A essas técnicas dá-se o nome de **aprendizado não-supervisionado**.

Neste trabalho utilizaremos dois métodos de aprendizado não-supervisionado para agrupar diferentes tipos de pessoas, de forma que conseguimos prever e/ou classificá-las de forma a prever seus comportamentos futuros. Para isso, escolhemos um dataset de segmentação de clientes de um determinado banco na Índia. Nossas implementações podem ser encontradas em nosso repositório, e o *dataset* utilizado aqui.

# 2. Escolha do conjunto de dados

Visando escolher um conjunto que atendesse o desafio e dificuldades esperadas de um trabalho tal como esse, vasculhamos o Kaggle, site de competições, notebooks e conjuntos de dados para aprendizado de máquina. Selecionamos cinco conjuntos cujas explorações constam em nosso GitHub:

- *Unsupervised Learning on Country Data* (link);
- *Data Science Job Salaries* (link);
- *Trending YouTube Video Statistics and Comments* (link);
- *Spaceship Titanic* (link);
- *Bank Customer Segmentation* (link).

Embora seu domínio fosse de nosso interesse, o primeiro conjunto possui poucos dados e seus atributos aparentam estar todos em bom estado, ou seja, apresentariam pouco desafio.

O segundo, por sua vez, possui muitos atributos categóricos – o que, a depender do pré-processamento a ser realizado, poderia tornar o agrupamento não-supervisionado pelos métodos escolhidos menos proveitosos. Além disso, seu domínio era de pouco interesse; partimos para outro conjunto.

O terceiro conjunto aborda vídeos virais do YouTube, trazendo consigo algumas informações como seus títulos, quantidade de curtidas, comentários, etc. Percebemos rapidamente que ele se trata de uma tarefa de processamento de linguagem natural e, por isso, o abandonamos.

O quarto, fruto de uma competição dos próprios desenvolvedores do Kaggle, apresenta um conto sobre uma espaçonave que sofreu um acidente em pleno espaço sideral; seus atributos são interessantes; há desafios a serem superados no pré-processamento para que o máximo de informação seja preservada. Contudo, ao investigarmos o conjunto mais de perto, percebemos que ele não segue um comportamento de grupos. Técnicas de classificação são extremamente valiosas neste, mas agrupamento não-supervisionado, não.

O último conjunto apresenta um equilíbrio em suas características. Seu domínio é suficientemente interessante; constam muitos dados nele (mais de um milhão); possui desafios a serem superados no pré-processamento; e, de acordo com as *tags* usadas por quem o colocou no Kaggle, serve para agrupamento não-supervisionado. Por tais motivos, escolhemos esse conjunto para o trabalho.

## 3. Estudo do conjunto de dados

No *dataset* escolhido existem 9 atributos, sendo eles:

- *TransactionID*, o ID da transação;
- *CustomerID*, o ID do cliente;
- *CustomerDOB*, a data de nascimento do cliente;
- *CustGender*, o gênero do cliente;
- *CustLocation*, a cidade em que o cliente reside;
- *CustAccountBalance*, o saldo bancário do cliente;
- *TransactionDate*, o dia da transação;
- *TransactionTime*, o horário em que ocorreu a transação;
- *TransactionAmount (INR)*, o valor da transação.

Antes de prepararmos esses atributos para o agrupamento, precisamos entendê-los. Os dois primeiros (os IDs) servem para tornar única uma transação e o cliente que a realizou. *CustomerDOB* e *TransactionDate* são datas e, portanto, são *strings* no formato DD/MM/YY. *CustGender* são *strings* iguais a F, M, ou T, representando o gênero do cliente. *CustLocation* são *strings* com o nome da cidade em que o cliente reside, e possui um total de 9275 valores únicos diferentes. *CustAccountBalance* e *TransactionAmount (INR)* são *floats*. Por fim, *TransactionTime* é um valor inteiro. Para fins de esclarecimento, a Figura 1 traz as cinco primeiras entradas do conjunto.

| | TransactionID | CustomerID | CustomerDOB | CustGender | CustLocation | CustAccountBalance | TransactionDate | TransactionTime | TransactionAmount (INR) |
|---|---|---|---|---|---|---|---|---|---|
| 0 | T1 | C5841053 | 10/1/94 | F | JAMSHEDPUR | 17819.05 | 2/8/16 | 143207 | 25.0 |
| 1 | T2 | C2142763 | 4/4/57 | M | JHAJJAR | 2270.69 | 2/8/16 | 141858 | 27999.0 |
| 2 | T3 | C4417068 | 26/11/96 | F | MUMBAI | 17874.44 | 2/8/16 | 142712 | 459.0 |
| 3 | T4 | C5342380 | 14/9/73 | F | MUMBAI | 866503.21 | 2/8/16 | 142714 | 2060.0 |
| 4 | T5 | C9031234 | 24/3/88 | F | NAVI MUMBAI | 6714.43 | 2/8/16 | 181156 | 1762.5 |

Figura 1: cinco primeiras entradas do conjunto

## 3.a. Pré-processamento

Os atributos que representam IDs foram imediatamente eliminados do conjunto. Eles representam informações importantes para os bancos envolvidos na transação, mas para os algoritmos de agrupamento utilizados, não.

*CustAccountBalance* e *TransactionAmount (INR)* foram mantidos exatamente como estão. Esses atributos já são numéricos e, para o agrupamento, basta aplicar uma normalização neles – tarefa que será realizada mais tarde.

*CustGender* possui três valores: F, M e T. Analisando mais a fundo esses valores, vemos que apenas um caso consta como T. Sob o risco de binarizar um espectro complexo da existência humana mas visando o ganho de simplicidade na visualização gráfica dos atributos e suas relações, eliminamos a entrada que contém *CustGender* igual a T. Assim, este atributo torna-se basicamente binário; convertemos os valores M para 0,0 e os valores F para 1,0.

*CustLocation* levanta uma problemática. São 9275 cidades diferentes neste atributo. Realizar um *one-hot encoding* poluiria nosso conjunto e convidaria a maldição da dimensionalidade para qualquer agrupamento a ser realizado nele; um *ordinal encoding* adicionaria uma relação de ordem entre as cidades que pode não ser interessante. Ademais, um problema que só pudemos notar após tentar um agrupamento pela primeira vez é que, por limitações técnicas dos integrantes deste grupo, qualquer computador ao qual temos acesso para de funcionar ao agruparmos todos os mais de um milhão de dados do conjunto.

Devido aos fatores supracitados, um compromisso foi feito. Utilizamos apenas as entradas que pertencem às três cidades com mais exemplos – que são, em ordem decrescente, Mumbai, Bangalore e New Delhi – e eliminamos este atributo.

*CustomerDOB* recebeu uma conversão simples. Como todas as *strings* obedecem o formato mencionado anteriormente, as separamos pelo ‘/’ e utilizamos apenas o ano de nascimento para calcular a idade do cliente (novo atributo *CustomerAge*). Dois tratamentos foram realizados concomitantemente: algumas entradas possuíam 1800 como ano de nascimento – as eliminamos; e se o ano de nascimento $X$ (no formato YY) fosse maior que 22, consideramos 1900 + $X$, e no caso contrário 2000 + $X$.

*TransactionDate* e *TransactionTime* foram combinados. *TransactionTime* são valores inteiros que podem ter de 1 a 6 dígitos e os interpretamos com uma lei de formação simples, elucidada pelos exemplos a seguir. Seja $T \in$ *TransactionTime*:

- se $T$ = 143207, então $T$ significa 14:32:07;
- se $T$ = 43207, então $T$ significa 04:32:07;
- …
- se $T$ = 7, então $T$ significa 00:00:07.

Com essa interpretação e algumas manipulações aritméticas (vide repositório), separamos temporariamente as informações de hora, minuto e segundo que esse atributo escondia. *TransactionDate*, tal como *CustomerDOB*, obedece o formato DD/MM/YY. Utilizando manipulação de *string*, separamos os dias, meses e anos

desse atributo.

Para reduzir nossa contagem de 6 novos atributos para apenas 1 que armazena a mesma quantidade de informação, lançamos mão do *UNIX timestamp*: convertemos os valores de ano, mês, dia, hora, minuto e segundo para a quantidade de segundos desta data a partir de 01/01/1970. Armazenamos esses novos valores e eliminamos todos os atributos obsoletos.

A Figura 2 mostra as cinco primeiras entradas do conjunto tratado. Vale compará-la com a Figura 1 para perceber a redução e seleção realizada nos atributos. Com esses cinco atributos realizaremos o agrupamento.

| | CustGender | CustAccountBalance | TransactionAmount (INR) | CustomerAge | TransactionTimestamp |
|---|---|---|---|---|---|
| **2** | 1.0 | 17874.44 | 459.00 | 26 | 1.470159e+09 |
| **3** | 1.0 | 866503.21 | 2060.00 | 49 | 1.470159e+09 |
| **6** | 1.0 | 973.46 | 566.00 | 30 | 1.470170e+09 |
| **7** | 0.0 | 95075.54 | 148.00 | 40 | 1.470168e+09 |
| **9** | 0.0 | 4279.22 | 289.11 | 38 | 1.470177e+09 |

Figura 2: cinco primeiras entradas do conjunto tratado

Com tais tratamentos aplicados, nosso conjunto apresenta agora 251810 entradas de um valor inicial de 1048567.

## 3.b. Visualização

Porém, antes de aplicar os algoritmos de agrupamento em nosso conjunto tratado, visualizamos algumas informações e relações presentes nele. A Figura 3 traz a distribuição dos dados entre as três maiores cidades e o gasto médio das transações nas mesmas.

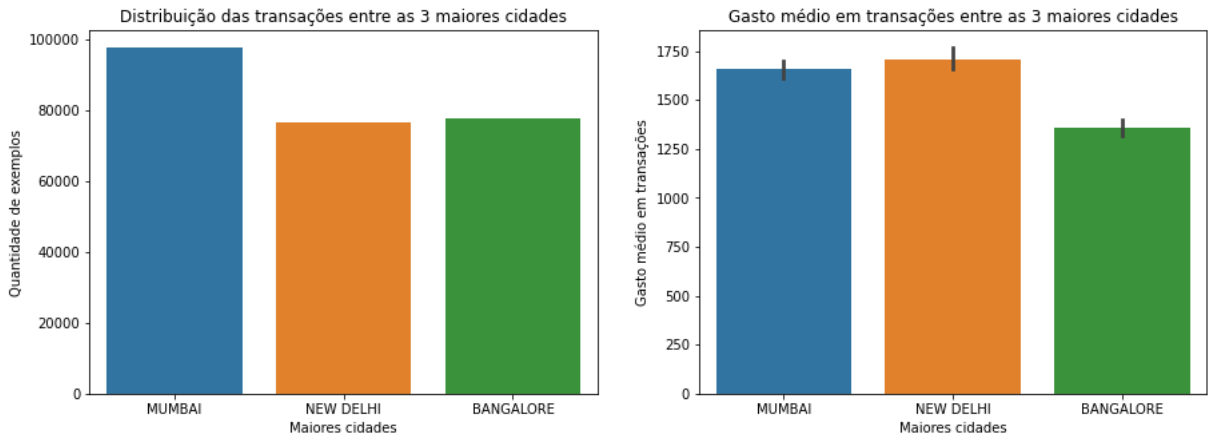



Figura 3: distribuição dos dados entre as três maiores cidades e o gasto médio das transações

A distribuição e frequência do saldo bancário nas transações pode ser visto na figura abaixo. De acordo com esse *violin plot*, percebe-se que esse conjunto é sintético,

i.e., não representa dados reais mas sim fabricados.

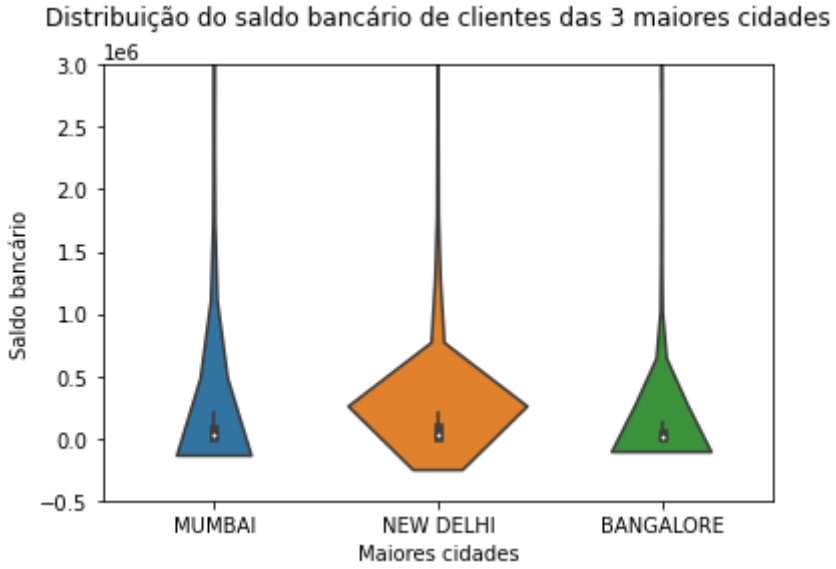



Figura 4: Distribuição do saldo bancário nas três maiores cidades

A Figura 5 apresenta a distribuição dos dados com relação aos meses do ano aos quais as transações pertencem. Traz, também, o gasto médio por transação em cada mês. Nota-se claramente que o mês de outubro possui pouquíssimas transações – de fato, realizando uma análise posterior, ele possui apenas 886 transações.

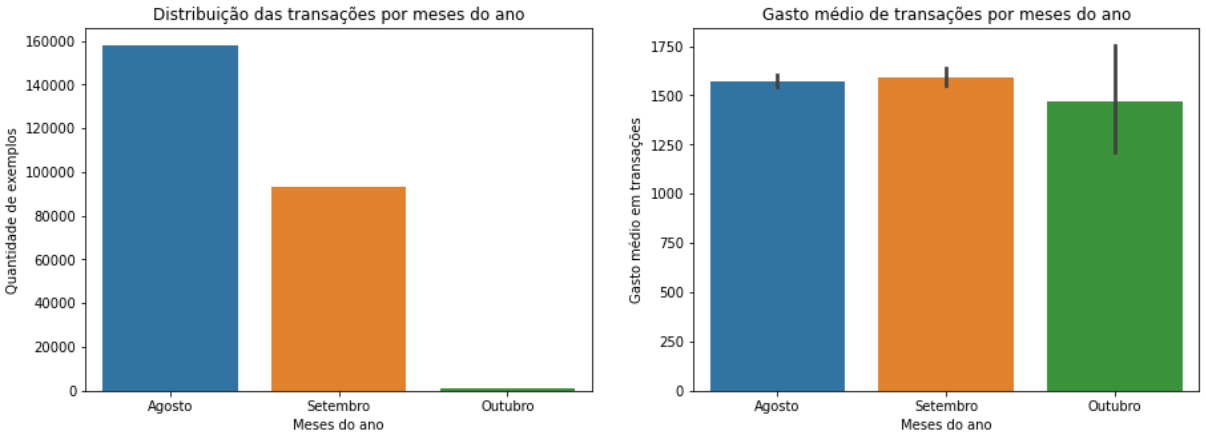



Figura 5: Distribuição dos dados entre meses do ano

O gasto médio entre os meses de agosto e setembro são praticamente iguais. Vejamos a forma com que esses gastos se desenvolvem ao longo dos dias de cada um desses meses. Ainda por cima, coloquemos uma separação: vamos ver os comportamentos médios de gasto entre homens e mulheres.

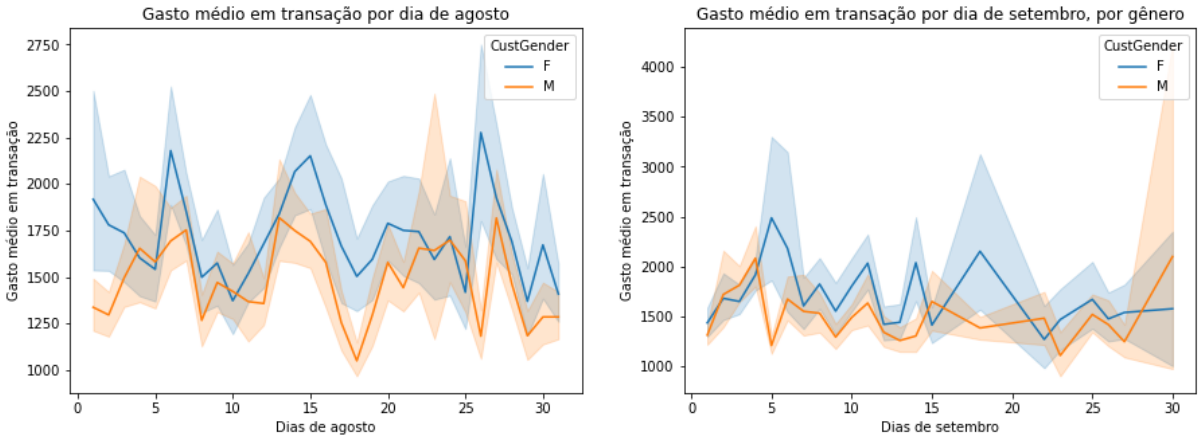



Figura 6: Gasto médio em transações pelos dias de cada mês (excluindo outubro)

Segundo os gráficos acima, mulheres tendem a gastar, em média, mais que os homens ao longo dos meses indicados. Ainda por cima, no mês de agosto é possível ver que três dias são os que mais possuem gastos: dia 6, dia 15 e dia 26.

Essa separação entre homens e mulheres foi proveitosa – mostrou claramente uma distinção no comportamento de cada um desses grupos. Analisemos essa separação mais um pouco.

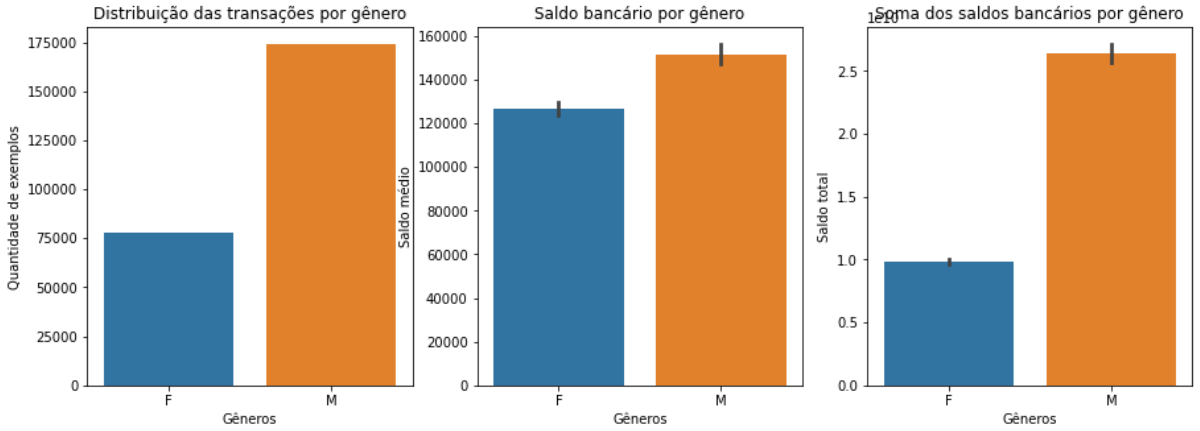



Figura 7: Diversas distribuições para os gêneros

Existem mais homens do que mulheres em nosso conjunto; o saldo médio dos dois gêneros é próximo, mas os homens concentram mais renda. Vejamos a distribuição, avaliando a partir do gênero do cliente de cada transação, entre os gastos por transação e o saldo bancário.

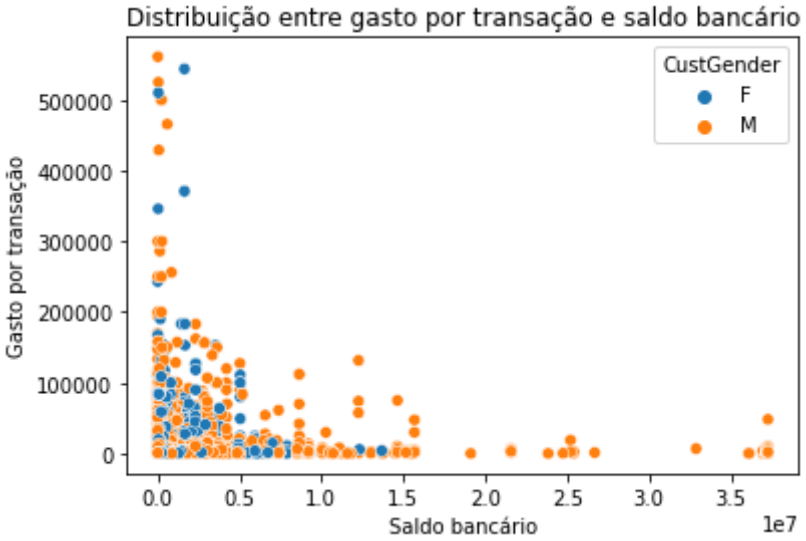



Figura 8: *scatterplot* entre gastos por transação e saldo bancário, por gênero

Não alcançamos nenhuma nova discussão sobre as separações dos gêneros com esse gráfico, mas é possível perceber um comportamento curioso nos nossos dados: à medida que as pessoas concentram mais renda, elas tendem a gastar menos em cada transação. Isso nos leva ao efeito contrário e contra-intuitivo: pessoas sem saldo bancário algum gastando centenas de milhares de rúpias indianas.

Após realizarmos os agrupamentos, mais análises serão feitas para ser possível interpretar os significados de cada grupo.

## 4. Algoritmos de agrupamento

Dado todas estas informações que conseguimos ver com o pré-processamento, aplicamos dois algoritmos de agrupamento, sendo eles o HDBSCAN (CAMPELLO; MOULAVI; SANDER, 2013) e o K-Means (MACQUEEN, 1967). Escolhemos tais algoritmos por conta de ser muito utilizado em várias áreas nos dias atuais.

## 4.a. HDBSCAN

Aplicando o algoritmo HDBSCAN com *alpha* igual a 1 e o tamanho mínimo do cluster a 600, foi retornado 4 grupos válidos e 1 grupo ruidoso (sendo o índice -1), sendo o tamanho deles:

| Índice do HDBSCAN | Quantidade de dados |
|---|---|
| -1 | 25341 |
| 0 | 77692 |
| 1 | 4956 |
| 2 | 1016 |
| 3 | 142805 |

Tabela 1: Saida do HDBSCAN.

Como o HDBSCAN tem um avaliador próprio que utiliza o Adjusted Rand Index juntamente com Overall F-measure. Assim, não fizemos uma avaliação utilizando outras métricas de qualidade.

## 4.b. K-Means

Para o algoritmo K-Means utilizamos duas métricas de avaliação, sendo elas o método do Cotovelo (THORNDIKE, 1953) e o Silhouette score (ROUSSEEUW, 1987).

Assim vemos nas figura 9, que o método do Cotovelo demonstra que se K estiver entre os valores 2 a 4, pode-se ter uma boa separação dos grupos, mesmo que a ideia do método seja considerar apenas uma queda brusca, utilizamos também o valor 3 e 4 para termos diferentes perspectivas do funcionamento e qual é o comportamento do mesmo. Porém a partir de 4, o erro quadrático mantém uma diferença muito pequena, a ponto de não ser tão relevante para a análise.

Para calcular o método do Cotovelo usamos a distorção para termos certeza de que ambos tivessem o mesmo valor.

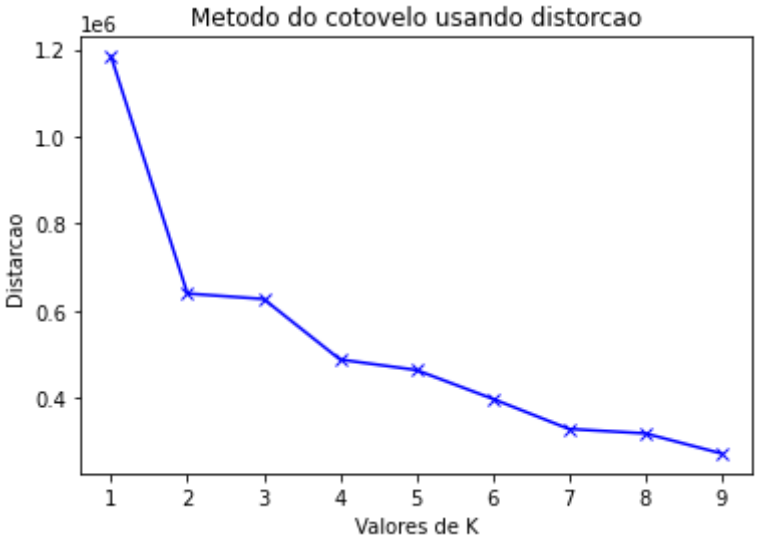



Figura 9: Gráfico do método do cotovelo.

O índice de Silhouette foi utilizado com valores entre 2 a 8, onde notamos que haveria análises mais interessantes para valores de K iguais a 2, 3 e 4. Segue na Figura 3 o resultado deste índice:

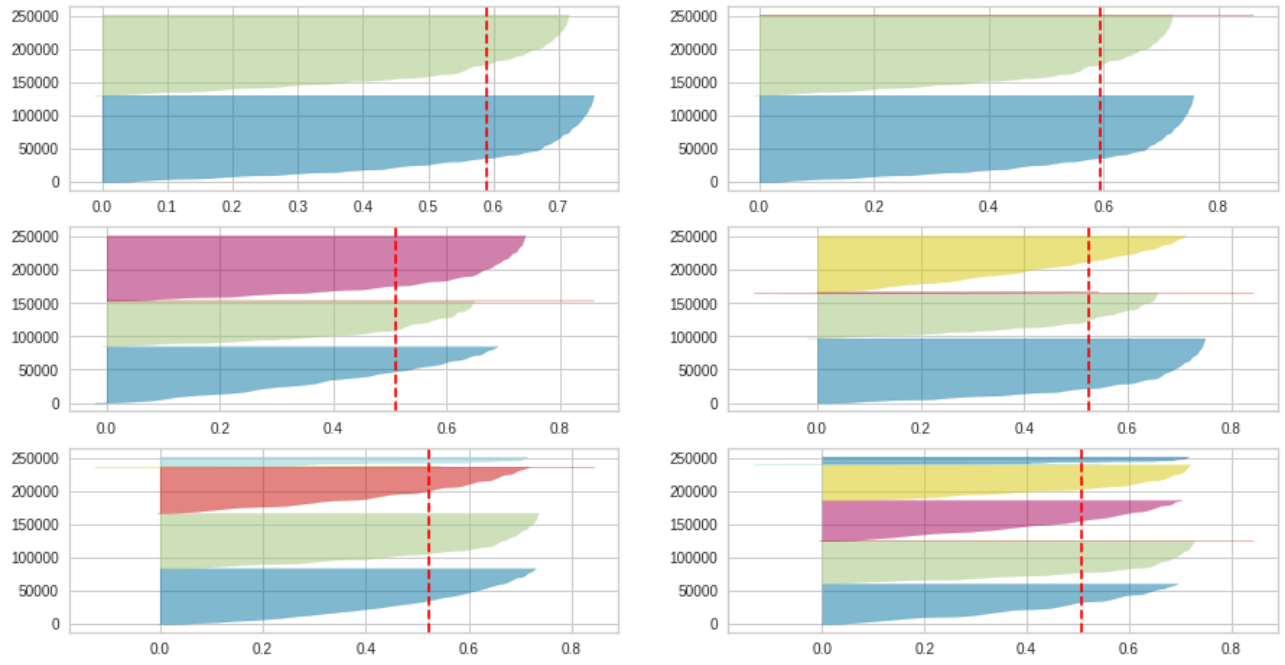


Figura 10: Gráfico de Silhouette com k entre 2 a 7.

Com esse gráfico conseguimos notar que a partir do valor 3, começa a aparecer ruídos sobre o gráfico, mesmo que ruídos pequenos e pouco notáveis, assim, fizemos também a análise do valor 3 e 4 sabendo que poderia ter valores ruidosos, porém, assim como mencionado anteriormente, queríamos verificar o funcionamento do K-Means mesmo com valores ruidosos.

# 5. Interpretação dos resultados obtidos

Portanto, combinando os resultados alcançados com o método do cotovelo e da silhueta, decidimos utilizar o K-Means com K variando de 2 a 4. O HBSCAN, tal como supracitado, já possui um avaliador próprio, então simplesmente o aplicamos com número mínimo de pontos em cada cluster igual a 600.

***K-Means (K = 2) –*** Analisando primeiro a distribuição das classes, percebe-se que a classe 0 possui quase o dobro de elementos do que a classe 1 (Figura 11). O saldo bancário médio e o gasto médio em transações das duas classes são parecidos, embora a classe 0 domine a primeira medida e a classe 1, a segunda (Figura 12).


Figura 11: Distribuição das classes (K = 2)

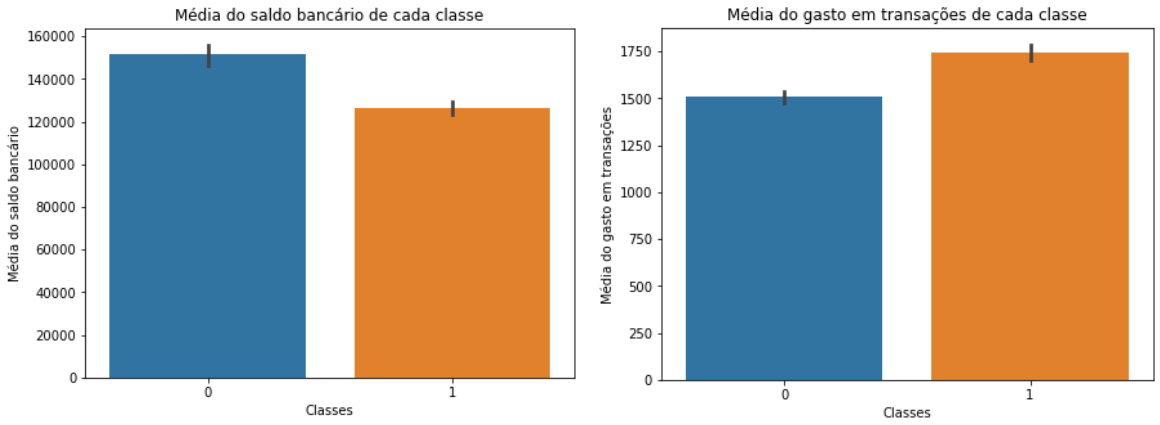
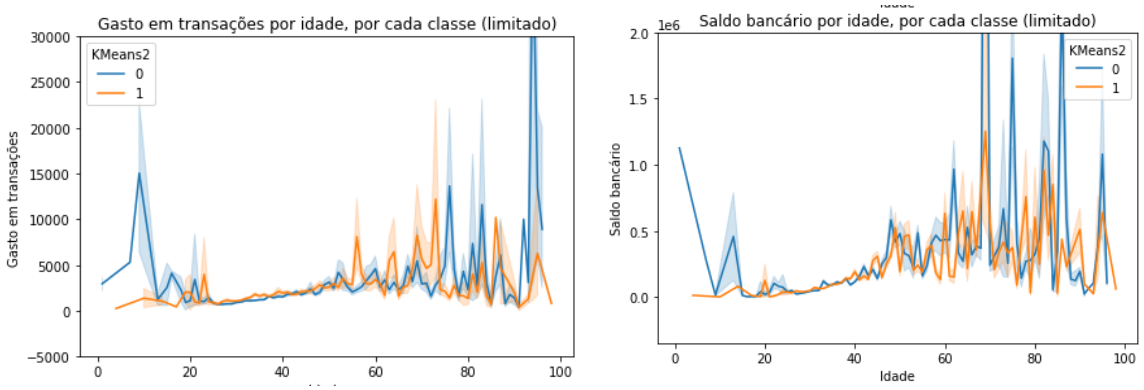



Figura 12: Saldo bancário médio e gasto médio em transação

Ao longo das idades presentes em cada uma dessas classes, conseguimos ver que, a não ser por alguns picos em que a classe 1 domina, a classe 0 tende a possuir maiores valores para gasto médio e saldo bancário médio (figura abaixo).




Porém, avaliando pelos dias do mês de agosto e setembro, outro comportamento é visto: a classe 1 gastou, em média, mais que a classe 0 (Figura 14).

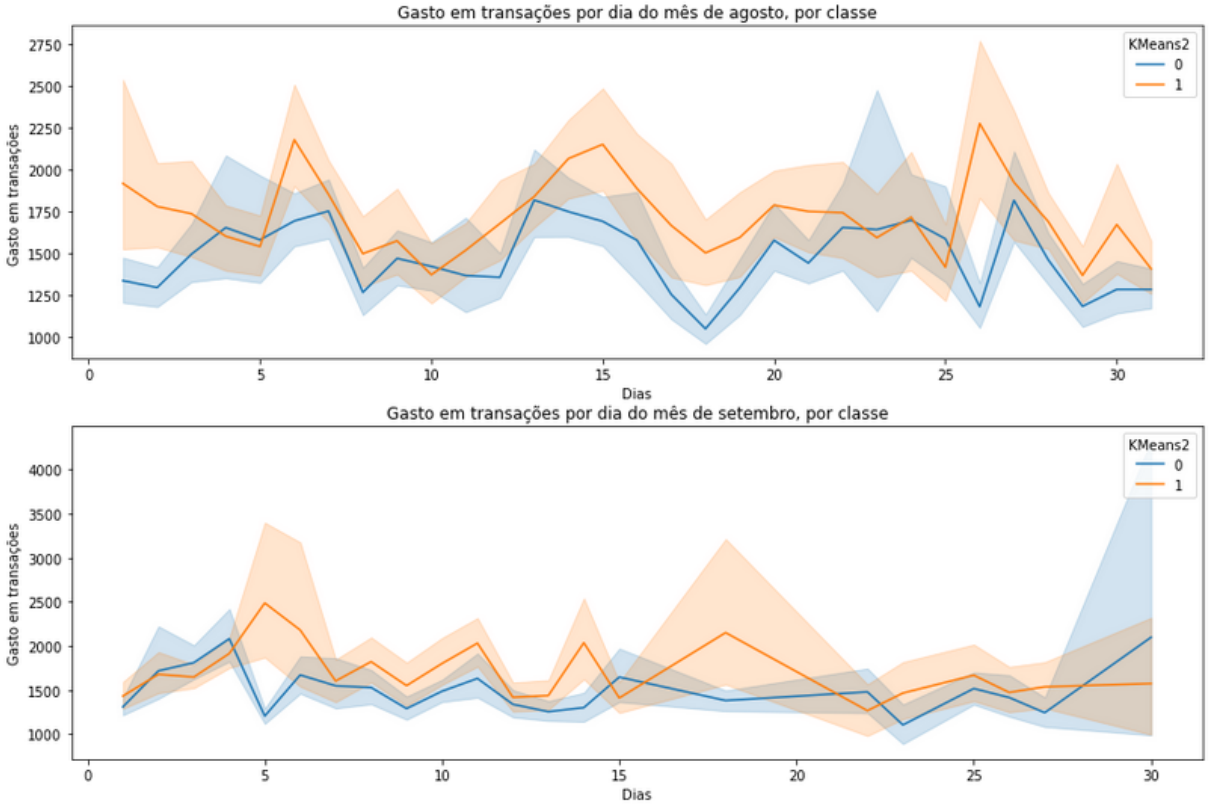



Figura 14: Distribuição dos gastos médios por dia de agosto e setembro

Poderíamos, então, assumir que a classe 0 trata de pessoas com uma concentração de renda sem igual pelas idades que possuem, mas a classe 1 tende a gastar mais em média. Isso indicaria dois tipos de transações: uma feita por concentradores de renda pouco dispendiosos e outra por pessoas sem saldo que gastam muito. Porém, a existência de mais um gráfico joga por água abaixo essa teoria; vejamos a Figura 15 logo abaixo:

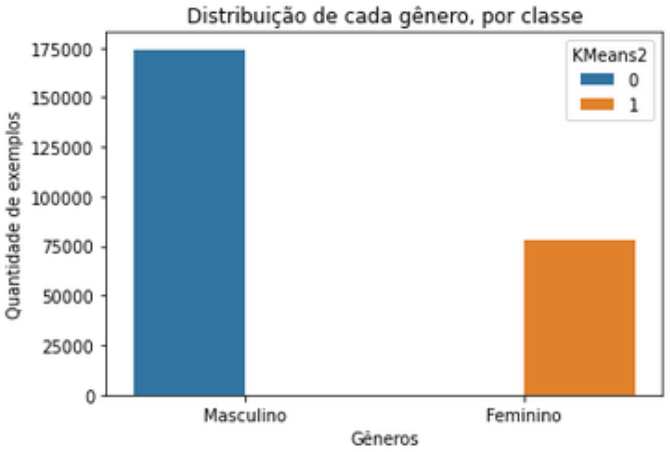



Figura 15: Distribuição das classes por gênero

De fato, se houvesse membros da classe 1 do gênero masculino, uma barra laranja acima de “Masculino” estaria presente. Mas não há. Para *K* = 2, nosso conjunto foi dividido perfeitamente entre homens e mulheres, cada qual com seus comportamentos financeiros. Não é difícil nos enganarmos por gráficos que aparentam indicar separação em tendências financeiras distintas, mas o agrupador não pensa como nós; dado que o atributo gênero é binário – 0,0 ou 1,0 –, ele separa claramente o conjunto em dois valores; todos os outros atributos possuem *ranges* em que suas entradas podem pertencer, mas gênero é um extremo ou outro extremo. Faz sentido esse ser o resultado do nosso agrupamento.

***K-Means (K = 3) –*** Comecemos dessa vez por gênero, então. A Figura 16 mostra a distribuição entre as classes para gênero.

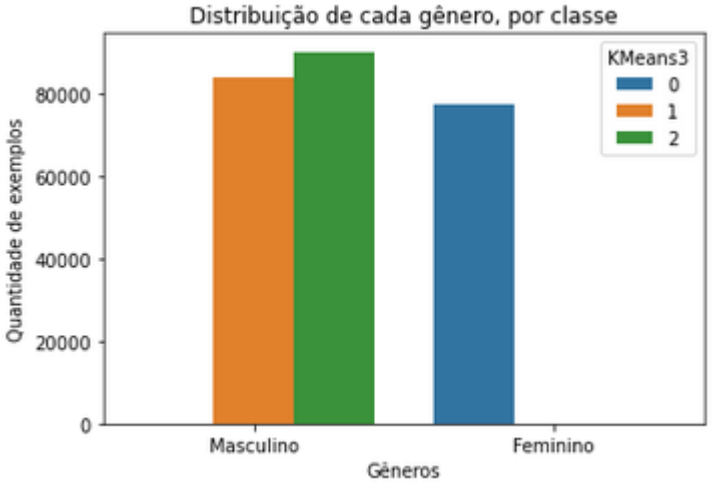



Figura 16: Distribuição de gênero entre as classes

Novamente, uma classe – dessa vez, a 0 – representa unicamente as mulheres do nosso conjunto. As classes 1 e 2 são partições dos homens; precisamos descobrir o que os particionou.

Avaliando o saldo bancário médio e o gasto médio em transação dessas classes, não vemos grandes diferenças (Figura 17).

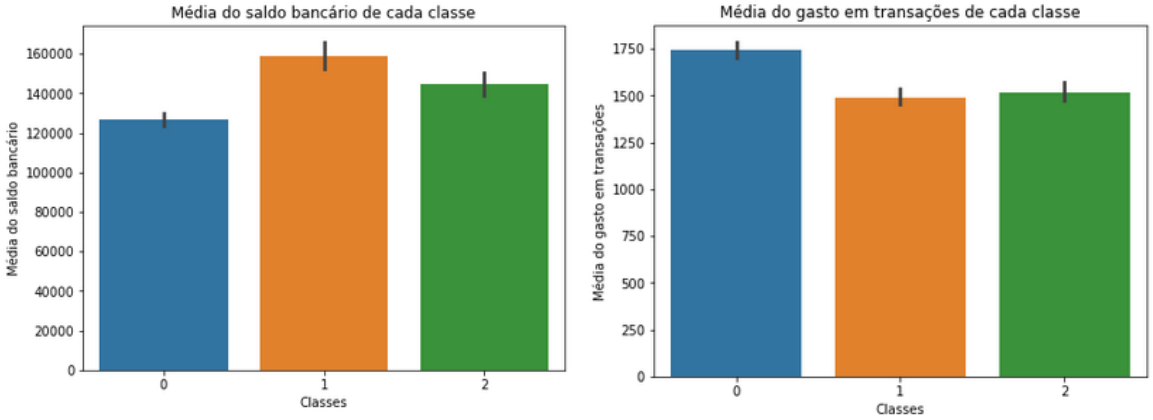



Figura 17: Saldo bancário médio e gasto médio por classe

Também não há grandes diferenças etárias. O verdadeiro ponto de cisão advém da análise dos momentos em que as transações foram realizadas. Vejamos a figura abaixo.

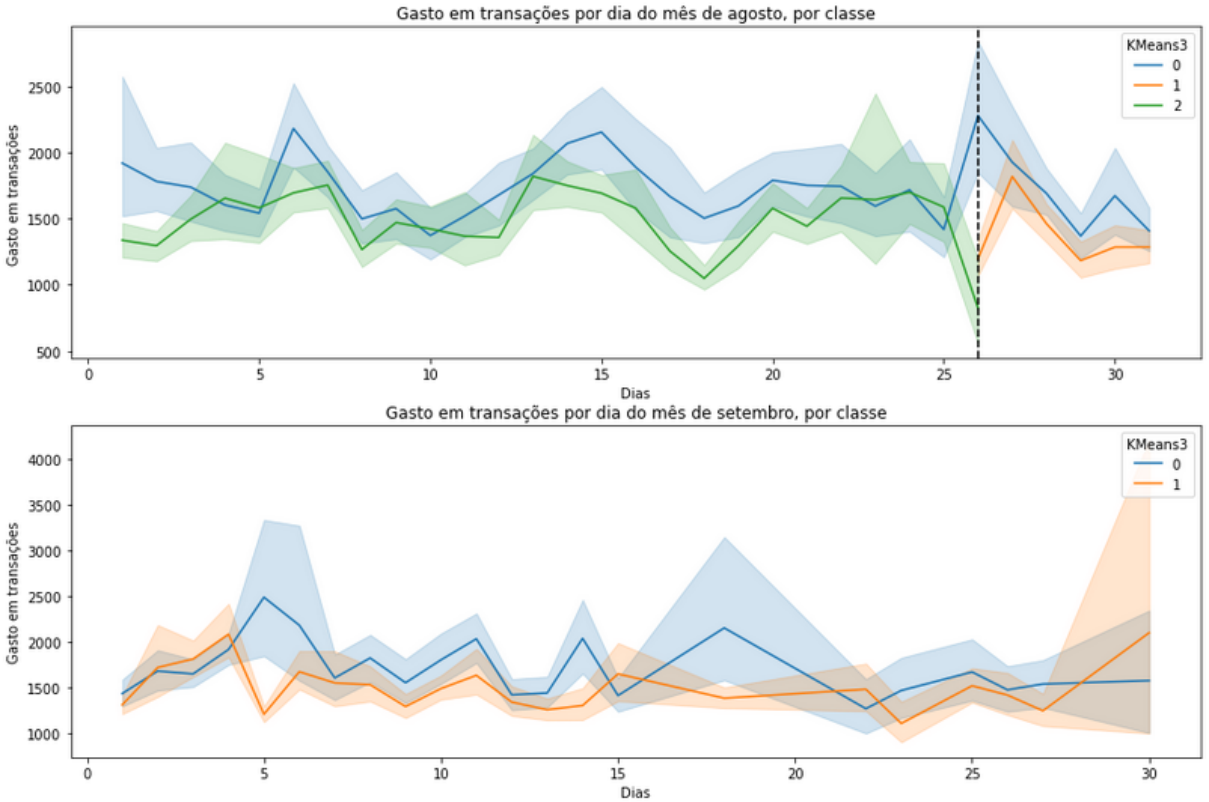



Figura 18: Distribuição dos gastos médios por dia de agosto e setembro

Há uma clara divisão: a classe 2 contém apenas transações realizadas antes de 26/08 (reta tracejada); após essa data, a classe 2 não possui mais nenhum membro. A partir desse gráfico conseguimos afirmar que, para $K$ = 3, a separação realizada

deu-se como *(mulheres) OU (homens antes do dia 26/08 OU homens depois do dia 26/08)*.

***K-Means (K = 4)*** **–** Para *K* = 3, houve uma divisão interessante levando em consideração o dia da transação. Para *K* = 2, a divisão deu-se no âmbito do gênero de quem realizou a transação. Levemos esses dois parâmetros em conta para avaliar o caso *K* = 4.

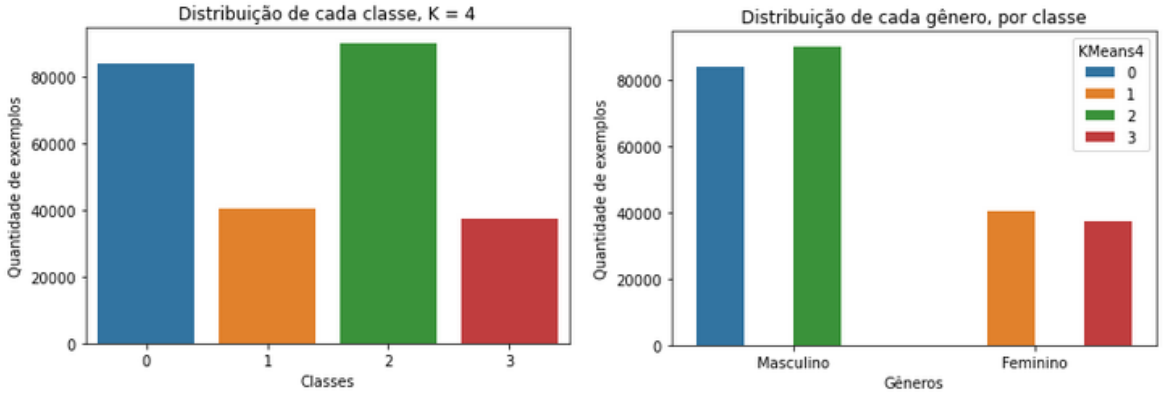



Figura 19: Distribuição de exemplos e gêneros por classe

A nova classe – classe 3 – possui apenas membros do gênero feminino. A cisão continua: classes 0 e 2 são representantes do gênero masculino, classes 1 e 3 do sexo feminino.

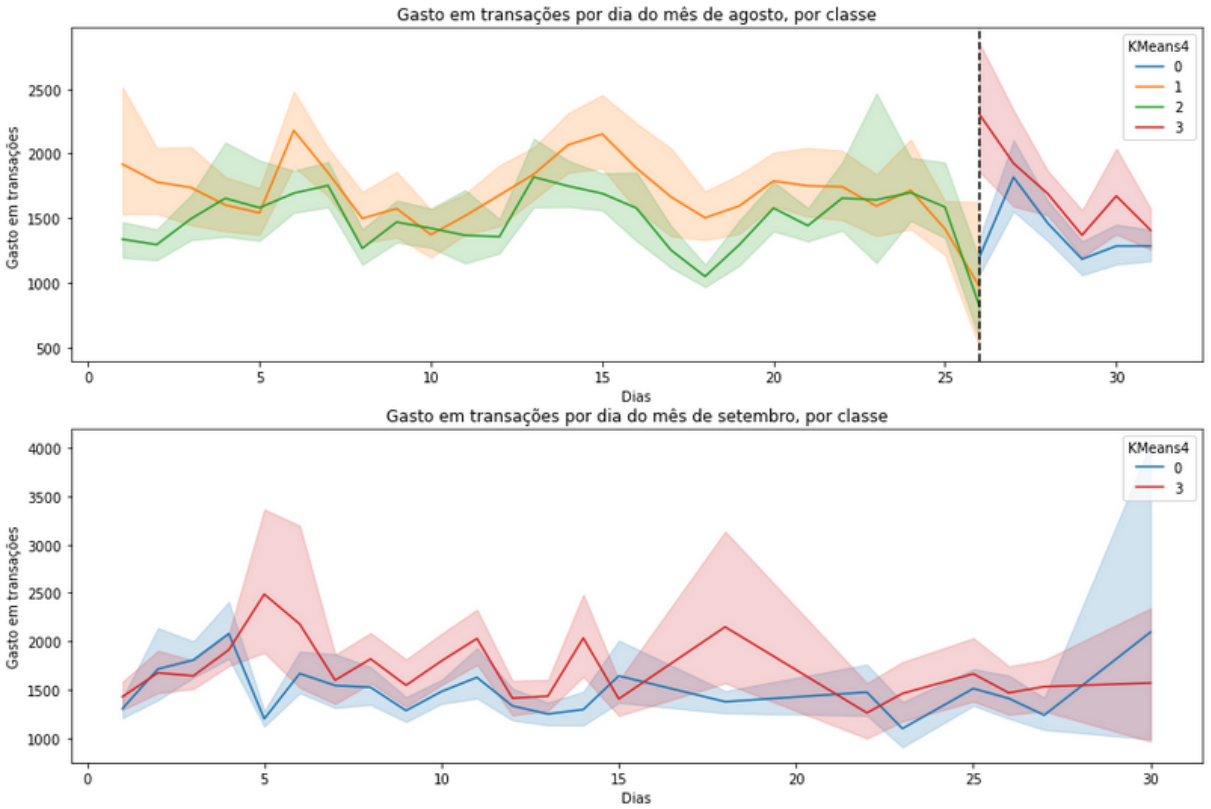



Figura 20: Distribuição dos gastos médios por dia de agosto e setembro

E neste ponto reside, novamente, a nova divisão. Tal qual para *K* = 3, nosso conjunto está sendo dividido entre homens e mulheres e transações realizadas antes e depois do dia 26/08. Mas dessa vez, essa diferença em data ataca também as mulheres.

As classes 1 e 2 representam transações realizadas por mulheres e homens, respectivamente, antes do dia 26/08; as classes 0 e 3 são transações realizadas por homens e mulheres, respectivamente, depois do dia 26/08.

***HDBSCAN –*** Assim como descrito acima, o HDBSCAN retornou 4 grupos válidos e um grupo de ruído, assim, descartamos esse grupo de ruído temporariamente, porém, ao final falaremos uma conclusão que chegamos sobre esse grupo.

Vemos na figura 21 que os grupos 0 e 3 tem uma densidade maior de pessoas enquanto o 1 e 2 mantém uma quantidade muito pequena de pessoas.

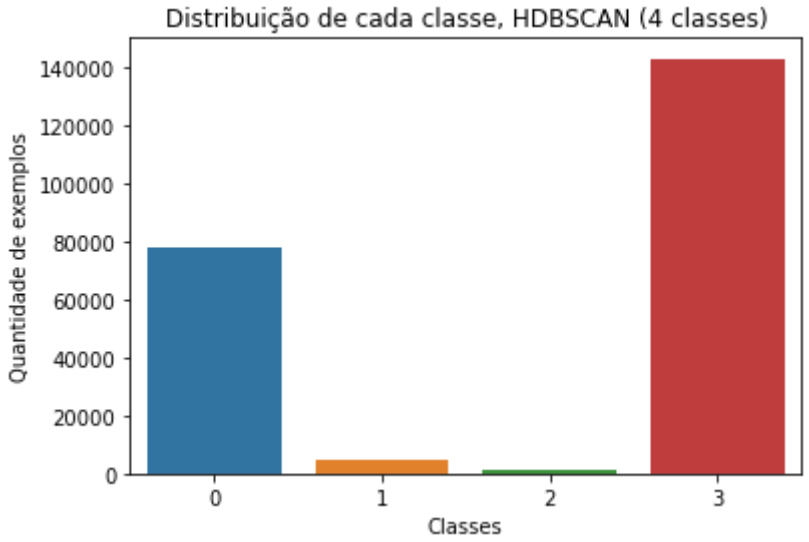



Figura 21: Distribuição do HDBSCAN.

Quando comparamos os saldos bancários e a soma dos mesmo por cada classe (figura 22) notamos que, a média dos saldos bancários da classe 2 e 3 são bem próximos, enquanto nos grupos 0 e 1 existe uma diferença grande. Porém, com isso ainda nada se conclui.

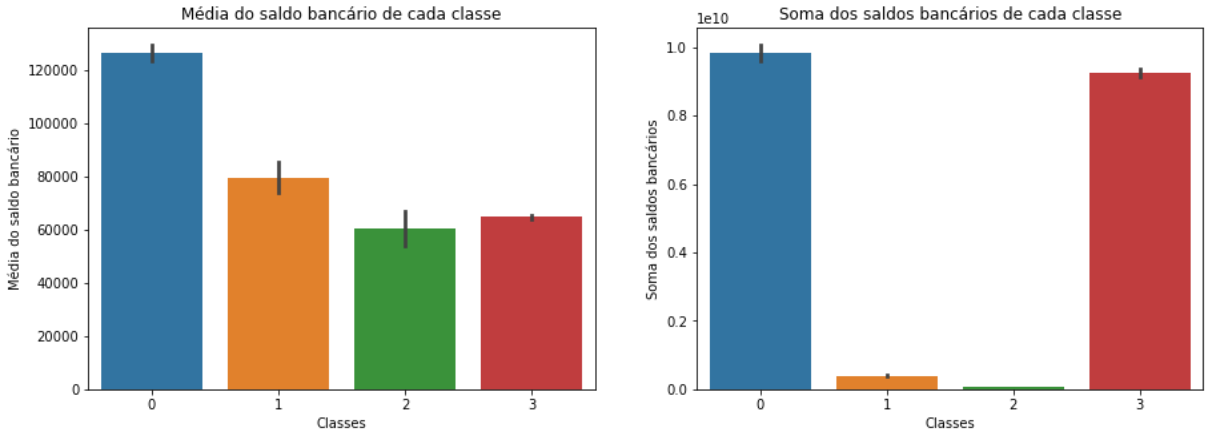



Figura 22: Gráficos de média de saldos bancários e a soma dos saldos bancários de cada grupo.

Agora analisando os gastos de cada classe (figura 23) conseguimos notar que o grupo 0 tem quantidades de gastos muito grande em compensação aos outros, porém como o grupo 3 tem muito mais pessoas, o total de seus gastos acabam sendo maior do que o grupo 0. Até então nada se conclui sobre os grupos 1 e 2, pois eles mantêm uma média de gastos próxima e as somas dos gastos são muito pequenas, provavelmente dado ao fato de que estes grupos são muito pequenos em consideração aos outros.

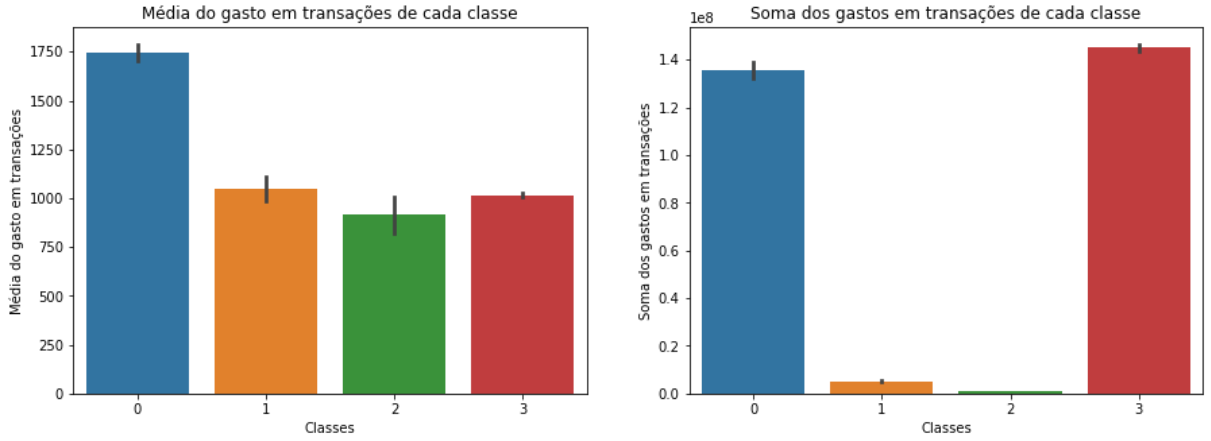



Figura 23: Gráficos de comparação da média e da soma total das transações de cada grupo.

Na figura 24, o gráfico demonstra a média etária entre cada classe, assim podemos julgar que a média de idade das pessoas são próximas ainda, ou seja, existe uma chance dos grupos 1 e 2 serem subgrupos apenas.

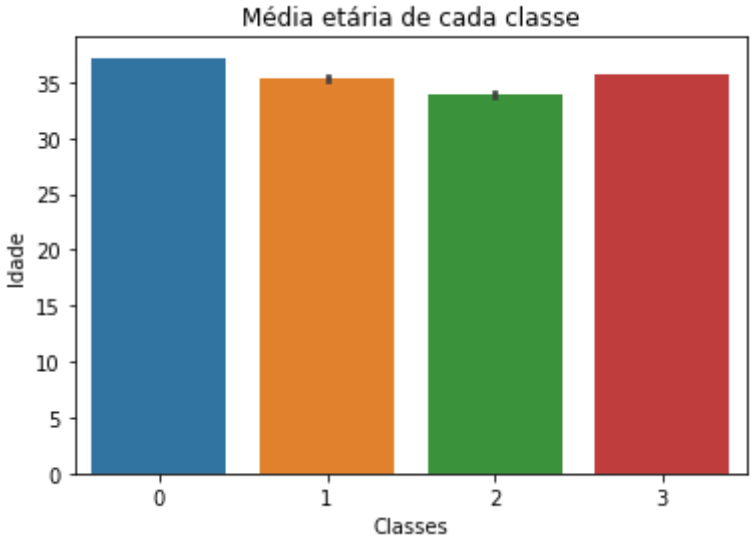



Figura 24: Gráfico da média etária de cada classe.

Na figura 25 é demonstrado o histograma dos grupos em consideração a idade e a quantidade de pessoas. Conseguimos notar nele que o grupo 0 tem pessoas de todas as idades enquanto os outros 3 grupo tem pessoas até 50 anos, dando a entender que os outros grupos podem ser subgrupos de fato, mas ainda não temos completa certeza sobre isso.

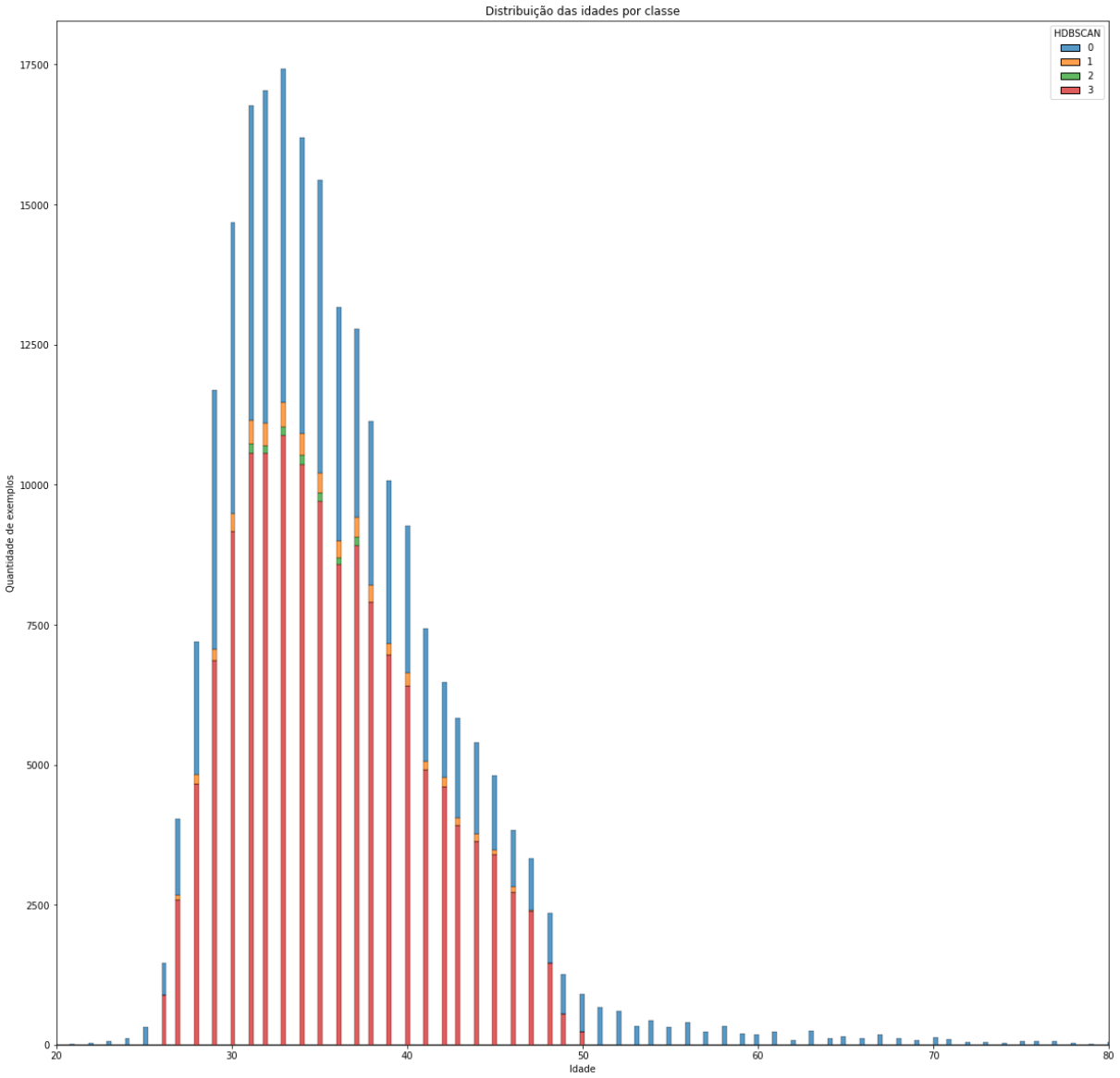



Figura 25: Histograma dos grupos.

Na figura 26 é demonstrado dois gráficos a direita e à esquerda são os mesmo gráficos só que demonstrados de forma limitada para termos uma melhor visualização. Analisando esses gráficos notamos que os grupos 1, 2 e 3 realmente são limitados a um intervalo próximo a 25 a 55, ainda não está claro o que o algoritmo decidiu como limiar para fazer tais separações, dado que um grupo é limitado a um intervalo do grupo 0.

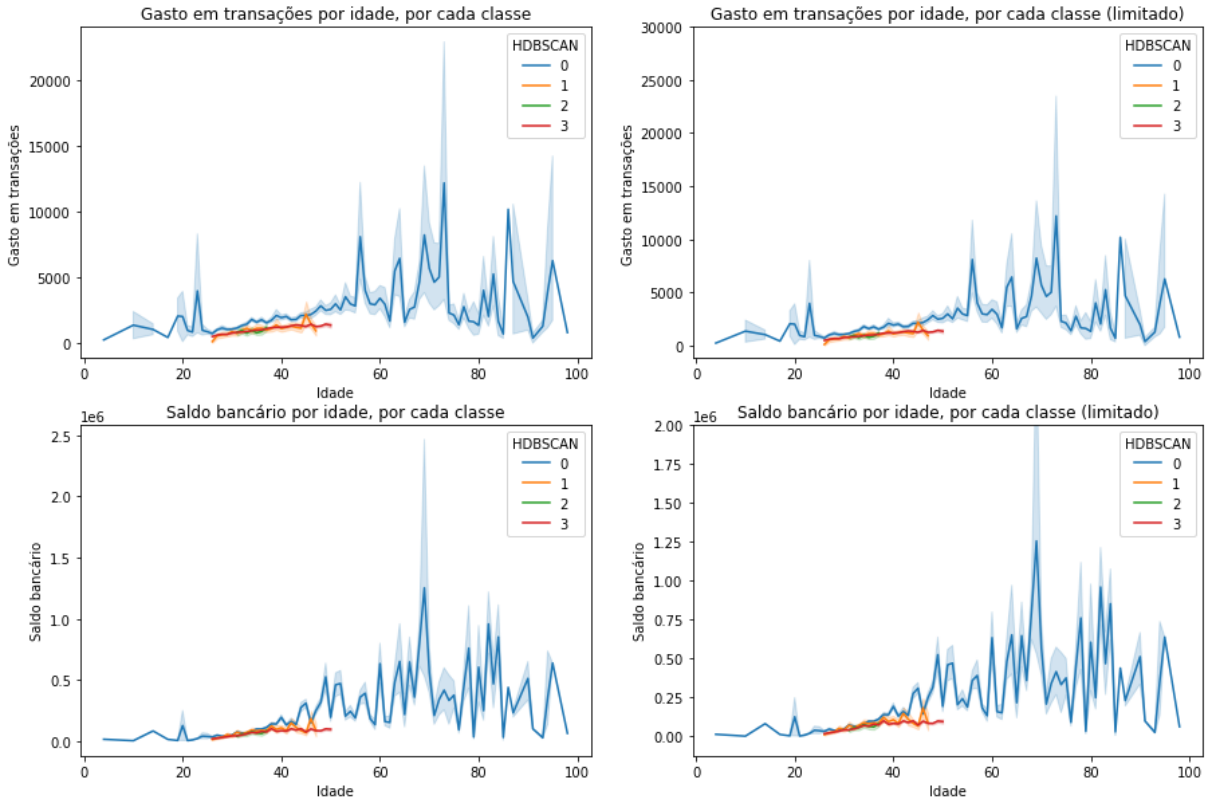



Figura 26: Gráfico de transações por idade e saldo bancário pela idade.

Na figura 27 à 29 é demonstrado a distribuição dos dados, a partir da 28 é reduzido um grupo para uma melhor visualização das distribuições.

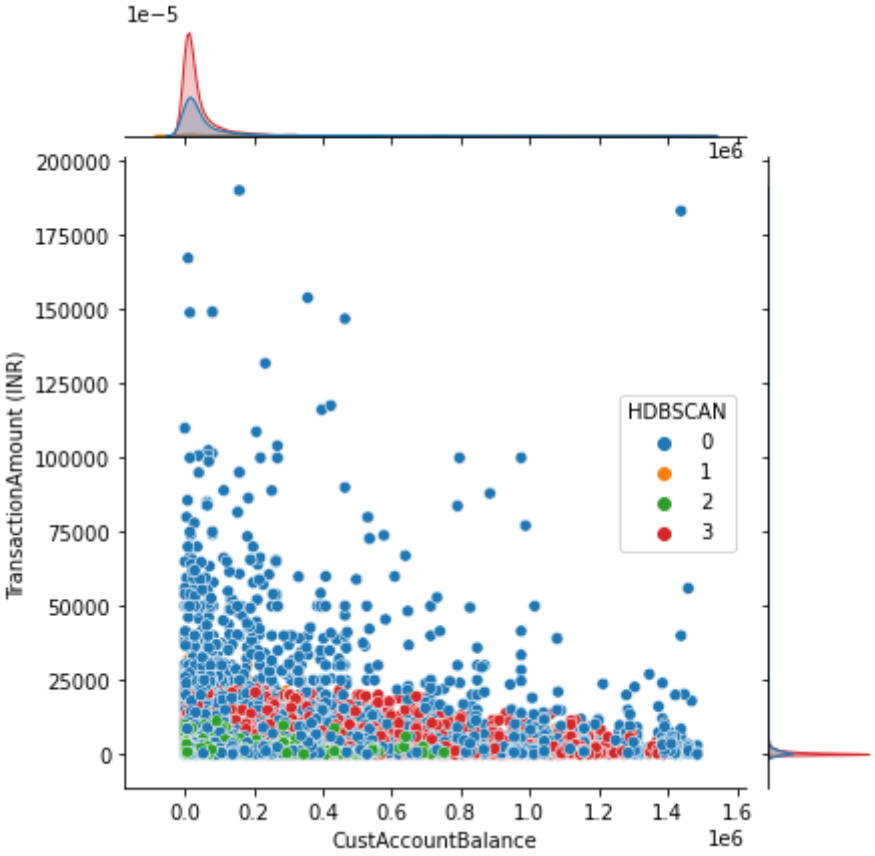



Figura 27: Distribuição dos dados e seus grupos.

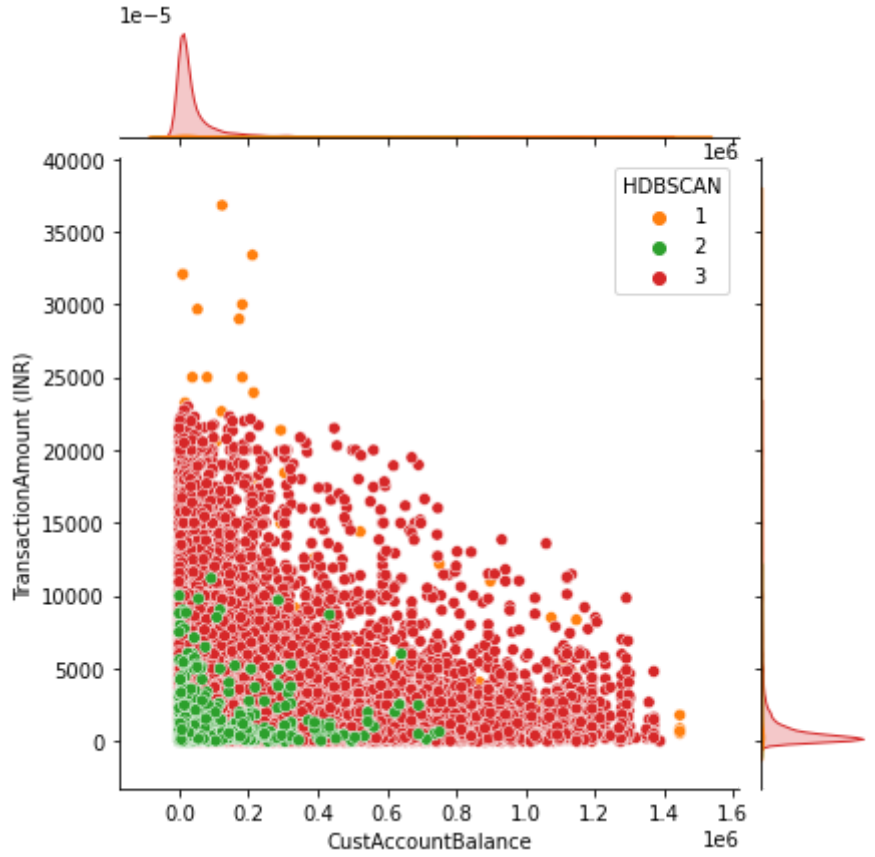


Figura 28: Distribuição dos dados com a remoção do grupo 0.

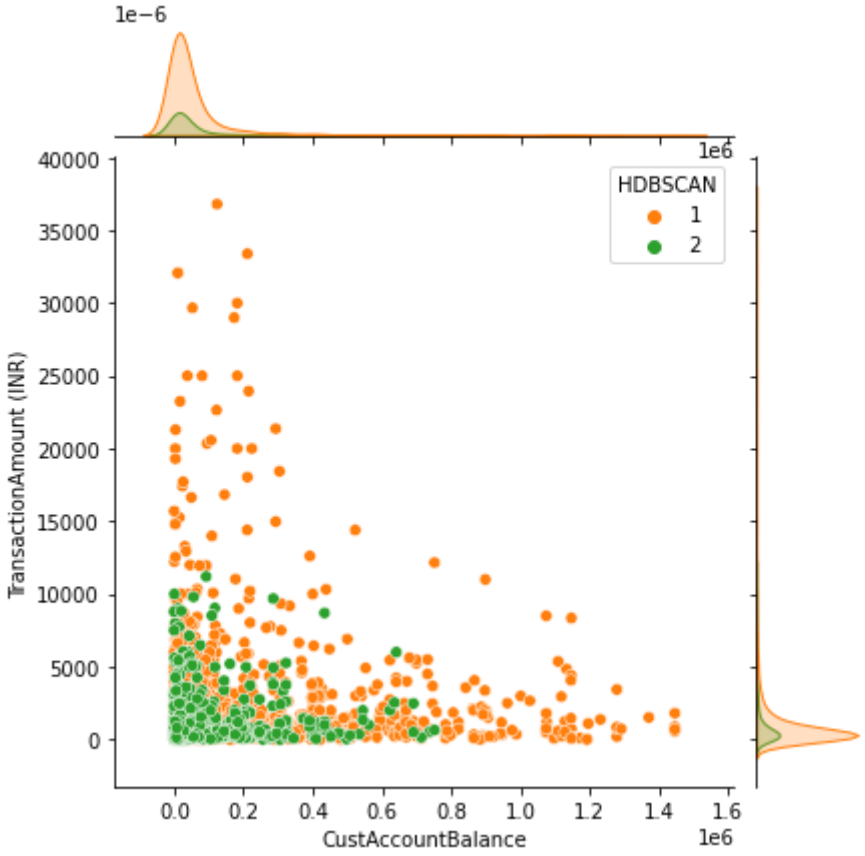



Figura 29: Distribuição dos dados com a remoção dos grupos 0 e 3.

Dado os gráficos demonstrados acima notamos que a distribuição dos dados se mantém bem diversa, porém não há um padrão ainda. Os grupos 0, 2 e 3 têm pessoas que fazem maiores transações e pessoas com maiores saldos nas contas, enquanto o grupo 1 tem, em sua maioria, pessoas com pouco dinheiro guardado na conta quanto poucas transações, dando a se entender que o grupo 1 é formado por pessoas com baixa renda provavelmente.

Já na figura 30, temos um comportamento mais interessante. Os grupos 1 e 2 fazem transações apenas em setembro e por um curto período de tempo, assim dando a se entender de que não se mantém uma linearidade do tempo que essas pessoas estão fazendo transferências, dando um destaque maior a teoria de que estes são subgrupos de um maior.

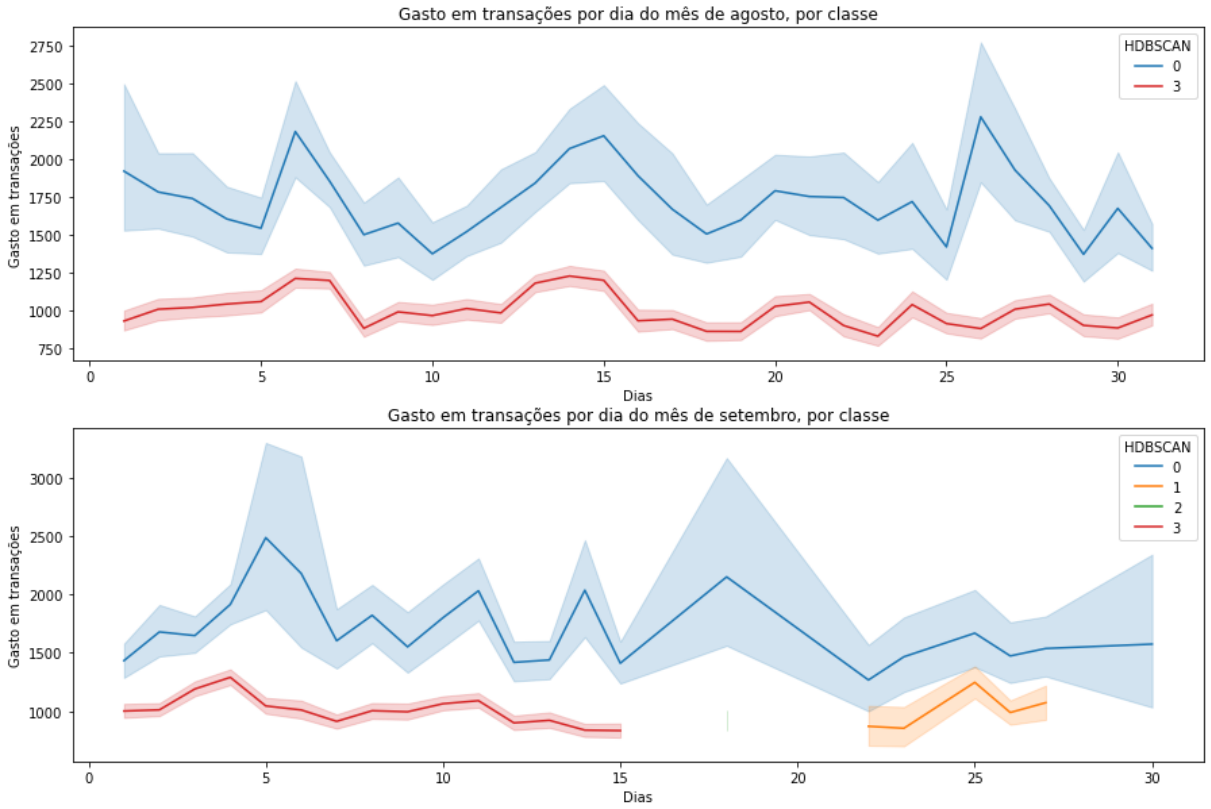



Figura 30: Gráfico de gastos em transações nos meses de agosto e setembro.

A figura 31 já nos confirma nossa teoria, os grupos 1 e 2 nos mostra que são do gênero masculino, enquanto o grupo 0, de maior quantidade de pessoas, são do gênero feminino.

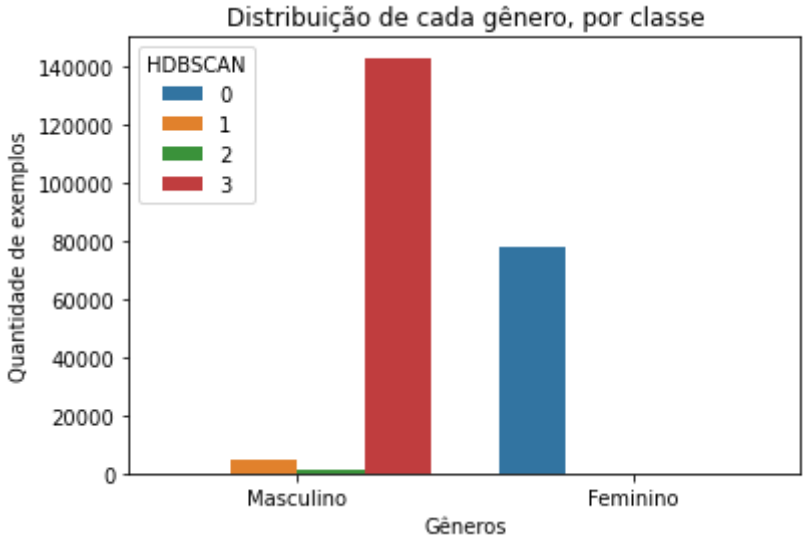



Figura 31: Distribuição de gênero por grupo.

Assim concluímos que, o HDBSCAN dividiu os grupos com a tendência maior a gênero, porém assim como falado anteriormente, as pessoas do gênero masculino foram divididos também considerando a classe social, dado que o grupo 2, tanto a quantidade de transações quanto o saldo são baixos.

# 6. Referências

CAMPELLO, Ricardo J. G. B.; MOULAVI, Davoud; SANDER, Joerg. Density-Based Clustering Based on Hierarchical Density Estimates. **Advances In Knowledge Discovery And Data Mining**, [S.L.], p. 160-172, 2013. Springer Berlin Heidelberg. http://dx.doi.org/10.1007/978-3-642-37456-2_14.

J. B. MacQueen, Some methods of classification and analysis of multivariate observations, In Proceedings 5th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, California, USA, 1967, 281–297.

ROUSSEEUW, Peter J.. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. **Journal Of Computational And Applied Mathematics**, [S.L.], v. 20, p. 53-65, nov. 1987. Elsevier BV. http://dx.doi.org/10.1016/0377-0427(87)90125-7.

THORNDIKE, Robert L.. Who belongs in the family? **Psychometrika**, [S.L.], v. 18, n. 4, p. 267-276, dez. 1953. Springer Science and Business Media LLC. http://dx.doi.org/10.1007/bf02289263.