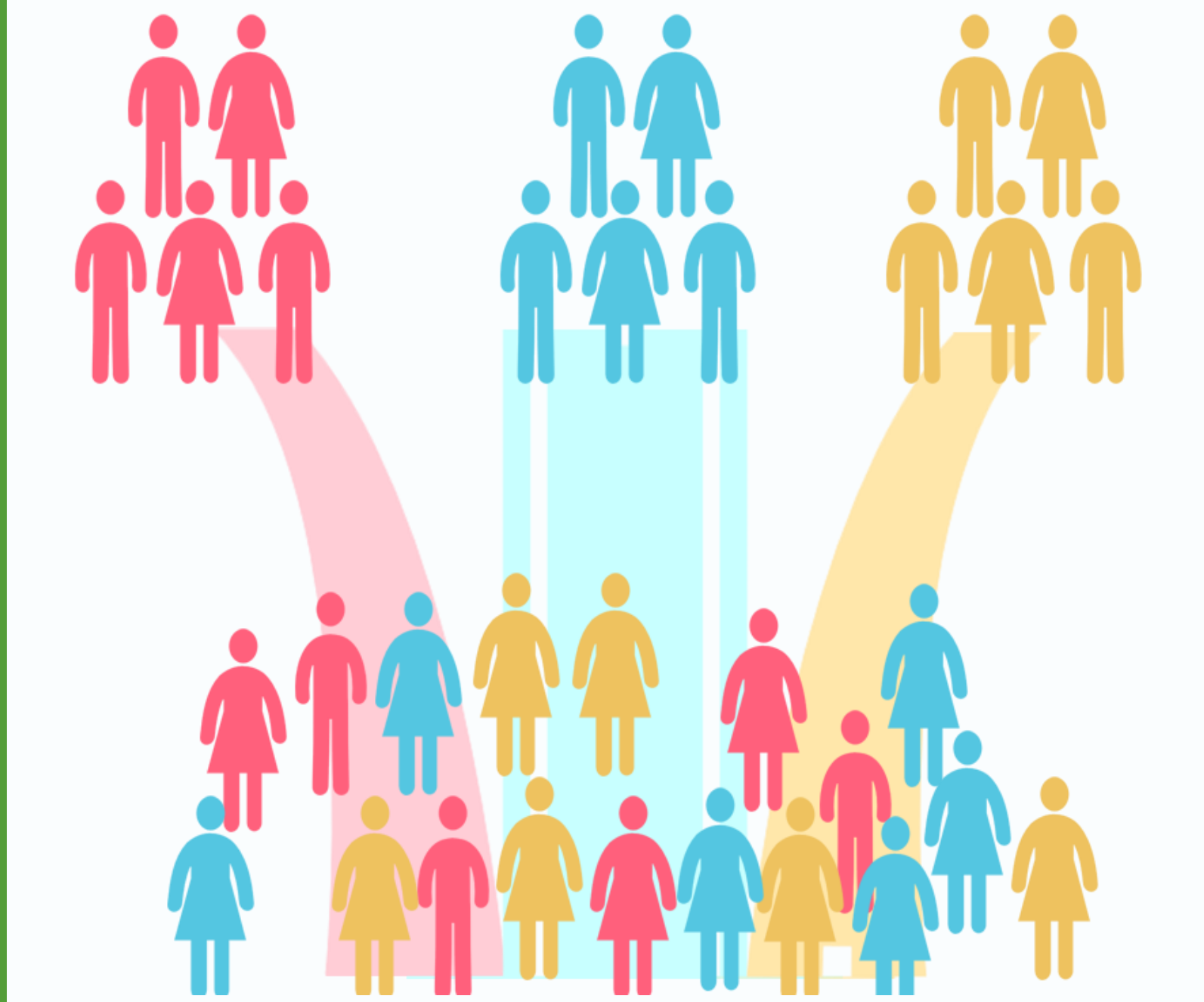
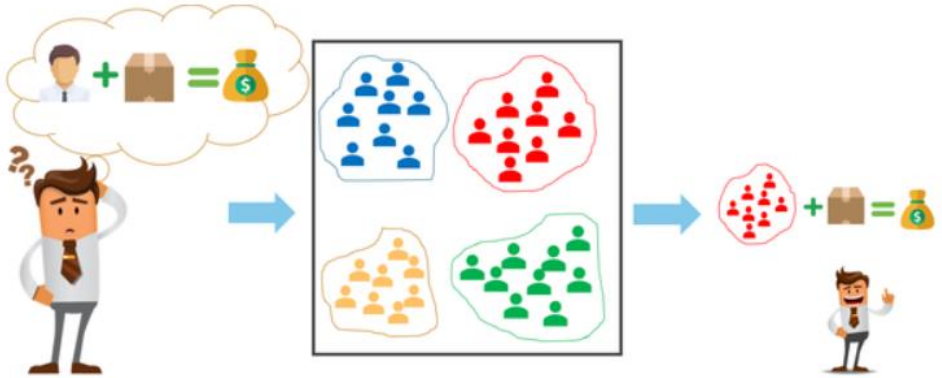
SEGMENTAÇÃO DE CLIENTES

PROJETO FINAL BOOTCAMP- INFNET



ALUNOS: MAXWELL MACIEL E VINICIUS BATISTA

INTRODUÇÃO



Neste trabalho faremos uma análise exploratória dos dados e aplicaremos algoritmos de Machine Learning não-supervisionados com o objetivo de segmentar os clientes de um shopping de acordo com sua similaridade, afim de identificar potenciais clientes, suas características e seus padrões de consumo.

Após a análise, o shopping poderá utilizar os resultados obtidos a fim de aumentar a assertividade do direcionamento das suas campanhas junto ao time de marketing, além promover eventos direcionados a fatia de clientes com maior potencial de compra para atraí-los. O banco de dados a ser analisado é de um período pré-pandemia do (covid-19).

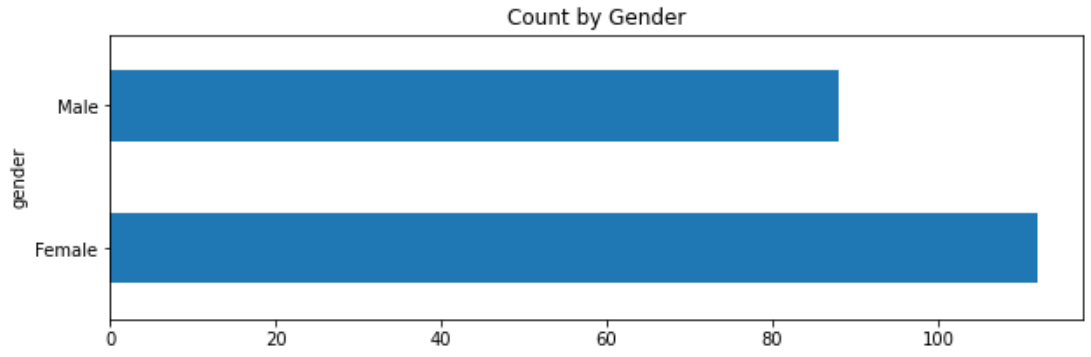
EXPLORAÇÃO DOS DADOS

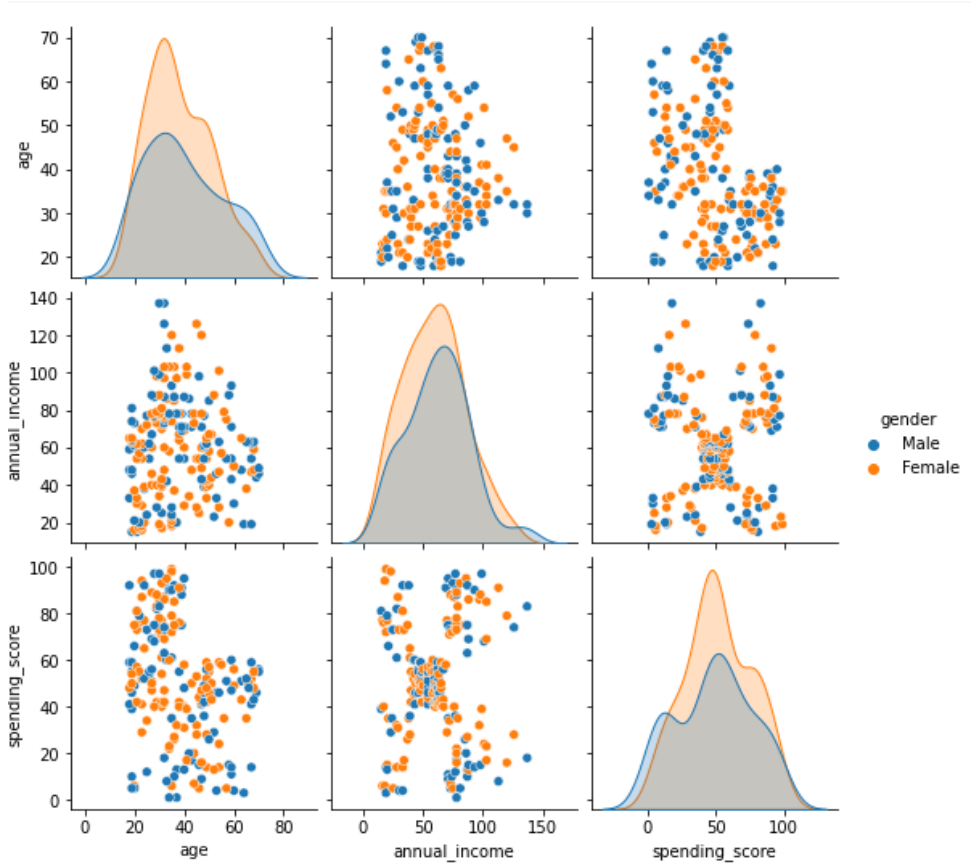
Carregando os dados

```
[ ]  df = pd.read_csv("https://raw.githubusercontent.com/MaxRodrigues91/INFNET_FinalProject/main/Mall_Customers.csv")
```

```
df.head(2)
```

	CustomerID	Gender	Age	Annual Income (k$)	Spending Score (1-100)
0	1	Male	19	15	39
1	2	Male	21	15	81

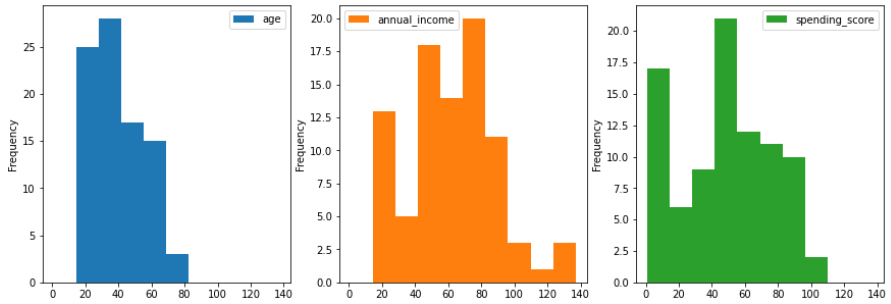




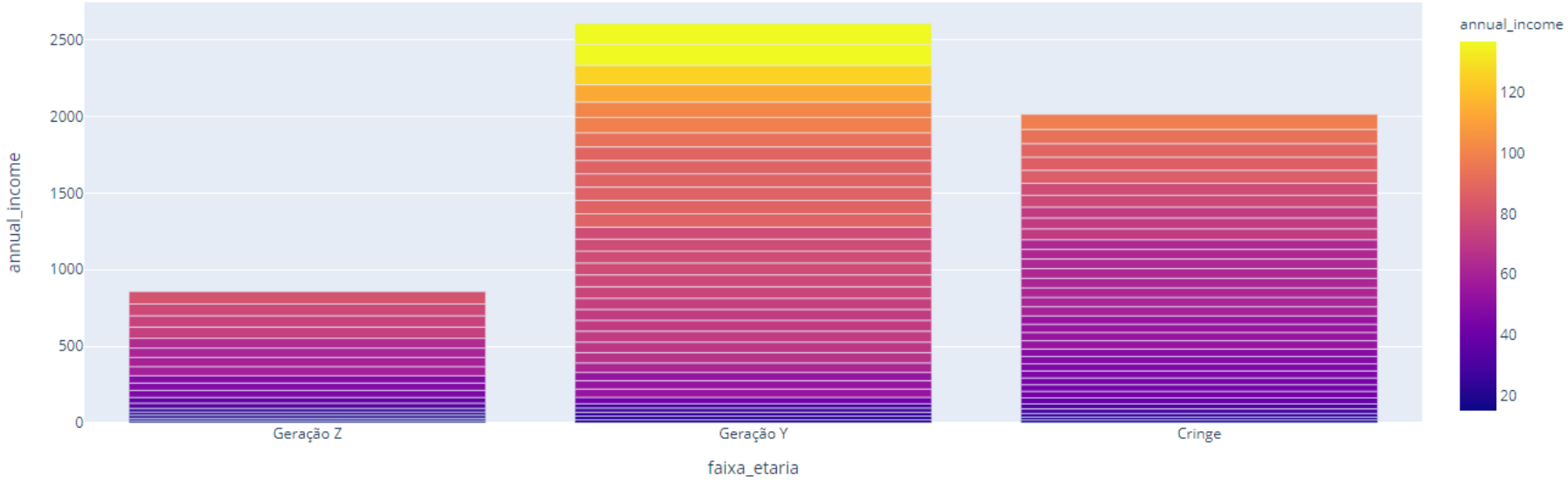
EXPLORAÇÃO DOS DADOS

Dados Gênero Masculino:

```
[ ]  fig,ax = plt.subplots(1,4,figsize=(20, 5))
     sdf2 = df1.groupby(by="gender")
     sdf2.get_group(1).plot(kind='hist',ax=ax, subplots=True, bins=10)
     plt.show()
```



Renda anual Clientes Gênero Masculino



EXPLORAÇÃO DOS DADOS

	spending_score								annual_income							
	count	mean	std	min	25%	50%	75%	max	count	mean	std	min	25%	50%	75%	max
faixa_etaria																
Cringe	35.0	34.942857	17.995004	3.0	16.0	41.0	48.50	60.0	35.0	57.514286	20.189356	19.0	45.00	60.0	71.00	98.0
Geração Y	35.0	60.885714	30.287613	1.0	51.0	69.0	85.50	97.0	35.0	74.457143	29.458189	20.0	58.00	77.0	87.00	137.0
Geração Z	18.0	50.833333	28.291030	5.0	39.5	53.5	71.25	92.0	18.0	47.611111	22.379452	15.0	26.25	48.0	63.25	81.0

Características do grupo do Gênero masculino

Cringe : Clientes que possuem um spending_score menor que 60 e possuem uma média de renda anual de 57,5(k).

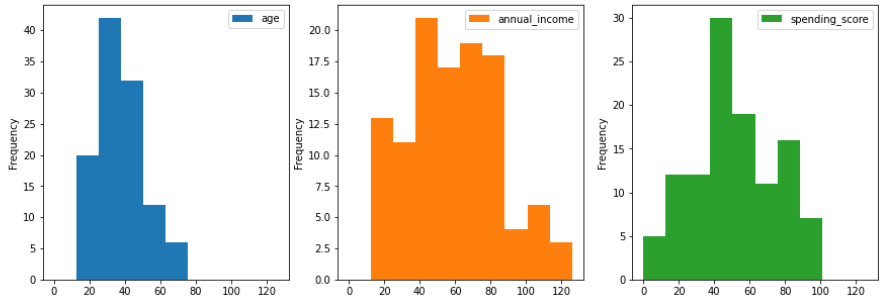
Geração Y : Clientes que possuem um spending_score de até que 97 (maior score do grupo masculino), é onde encontra-se o maior desvio padrão. Possuem uma média de renda anual de 57,5(k) e maior renda anual do grupo masculino.

Geração Z : Número de amostras 50% menor que os demais grupos.Possuem um spending_score de até que 92 (segundo maior do grupo masculino). Possuem a menor média de renda anual de 47,6(k) e a menor renda anual.

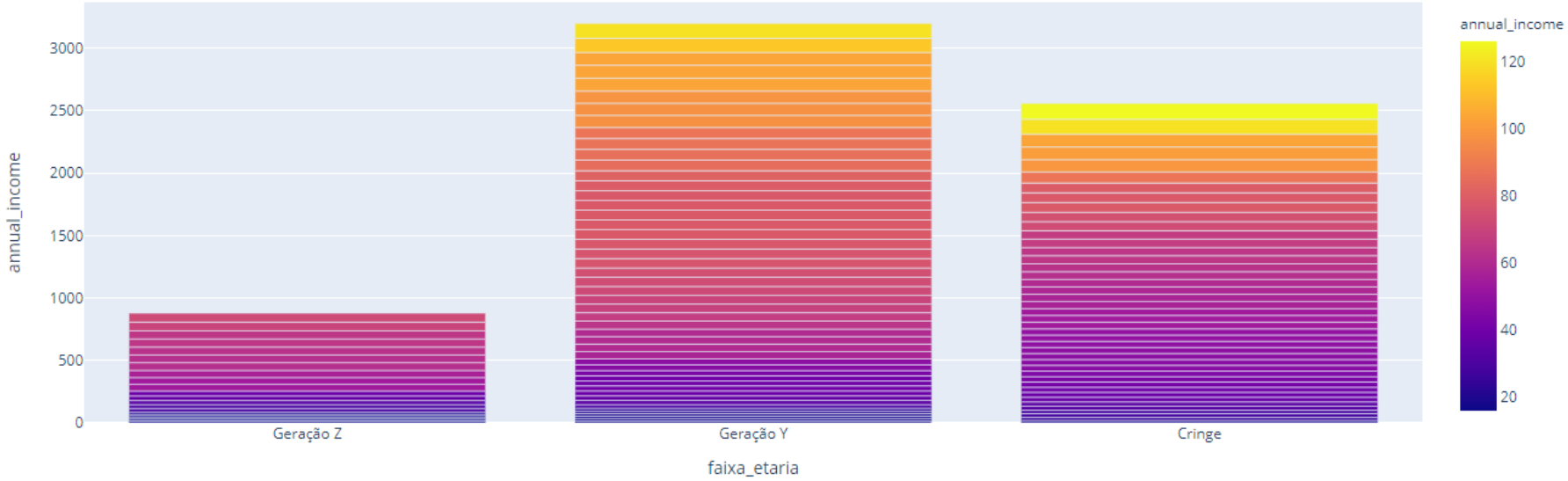
EXPLORAÇÃO DOS DADOS

Dados Gênero Feminino:

```
[ ]  fig,ax = plt.subplots(1,4,figsize=(20, 5))
     sdf2 = df1.groupby(by="gender")
     sdf2.get_group(0).plot(kind='hist',ax=ax, subplots=True, bins=10)
     plt.show()
```



Renda anual Clientes Gênero Feminino



EXPLORAÇÃO DOS DADOS

	spending_score								annual_income							
	count	mean	std	min	25%	50%	75%	max	count	mean	std	min	25%	50%	75%	max
faixa_etaria																
Cringe	43.0	36.813953	17.068846	5.0	18.5	43.0	50.00	59.0	43.0	59.488372	24.341198	20.0	43.50	54.0	70.00	126.0
Geração Y	49.0	61.530612	24.121310	6.0	42.0	69.0	83.00	99.0	49.0	65.285714	27.335112	17.0	40.00	72.0	81.00	120.0
Geração Z	20.0	58.650000	21.955158	6.0	46.5	56.0	76.25	94.0	20.0	43.950000	20.371484	16.0	26.75	46.5	62.25	72.0

Características do grupo do Gênero feminino

Cringe : Clientes que possuem um spending_score menor que 59 e possuem uma média de renda anual de 59,5(k) e a maior renda máxima anual.

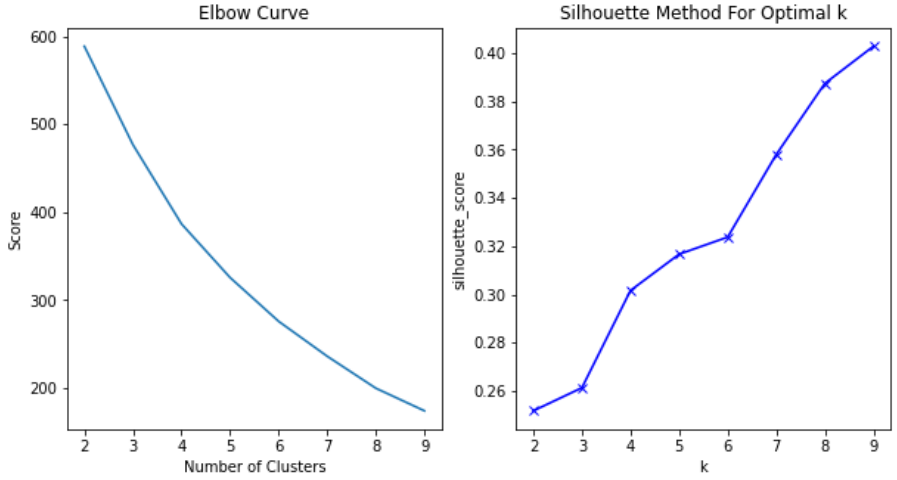
Geração Y : Clientes que possuem um spending_score de até que 99 (maior score do grupo feminino), é onde encontra-se o maior desvio padrão. Possuem uma média de renda anual de 65,3(k).

Geração Z : Número de amostras é aproximadamente 40% menor que os demais grupos. Possuem um spending_score de até 94 (segundo maior do grupo feminino). Possuem a menor média de renda anual de 43,9(k) e a menor renda anual.

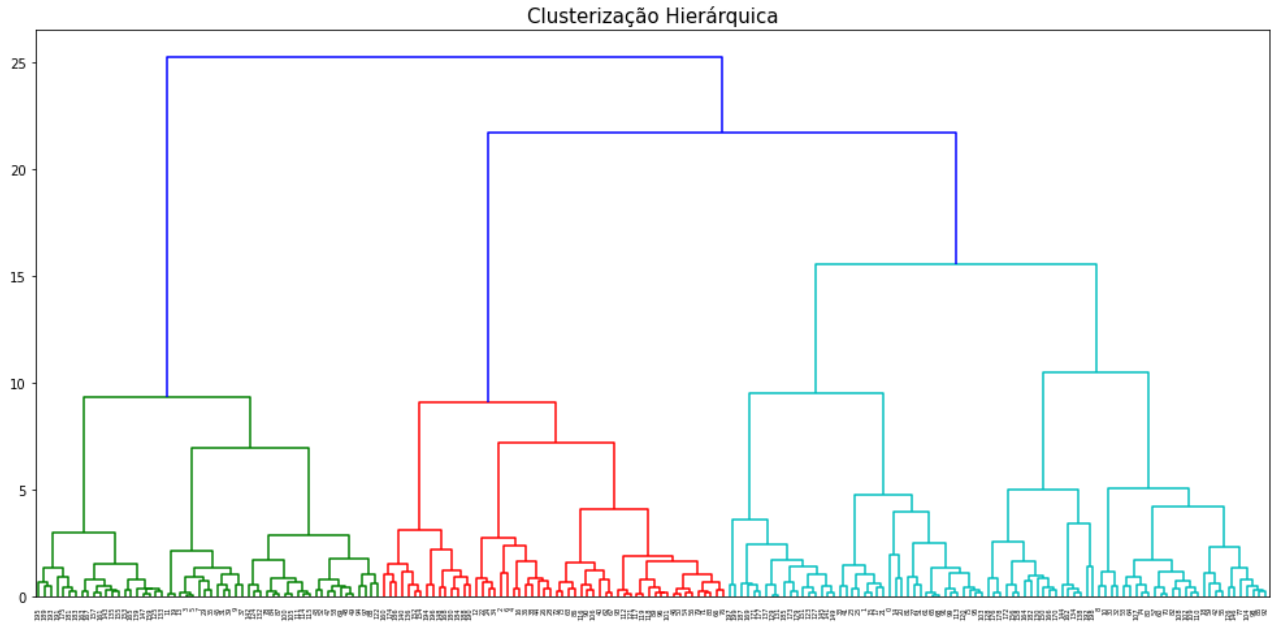
METODOLODIA

Foi realizada uma pesquisa prévia em estudos científicos sobre os melhores algoritmos para a solução do problema em questão. O k-means apareceu como o principal algoritmo em estudos de segmentação de clientes, tanto se tratando do número de vezes utilizado, como na qualidade dos resultados gerados.
Seguido de Análise de Componente Principal(PCA) e Clusterização Hierárquica (HC) .
Em nossa primeira análise iremos normalizar os dados e clusterizar com todas as entradas para verificar qual comportamento teremos.

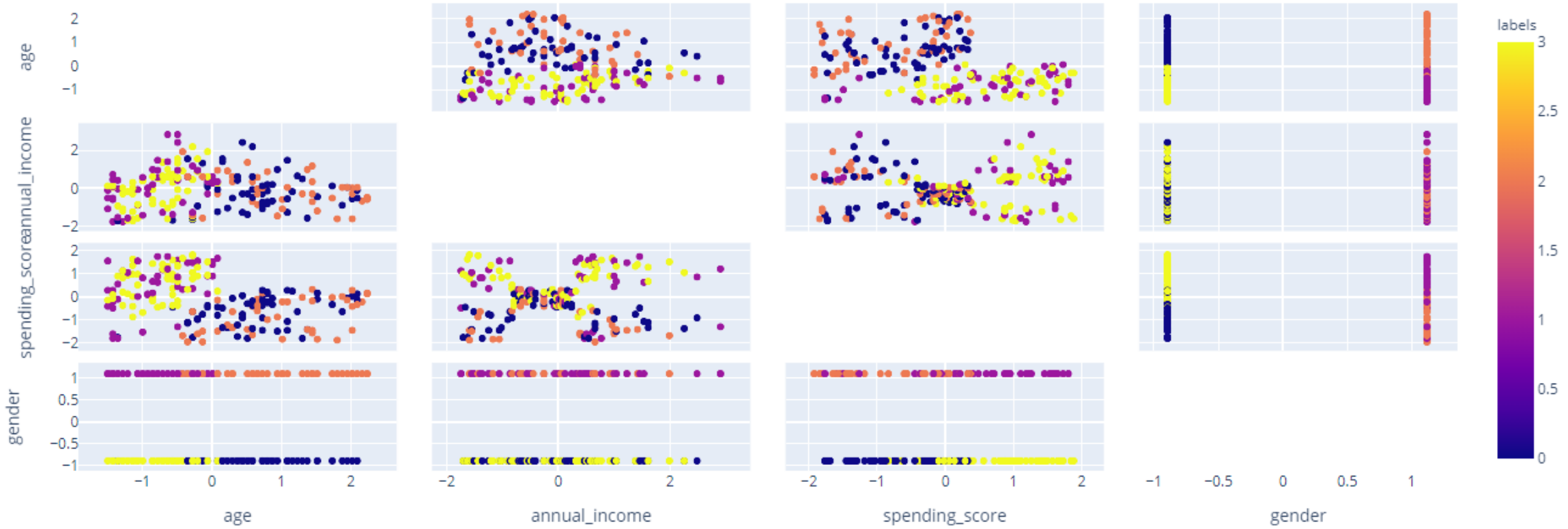
K-MEANS



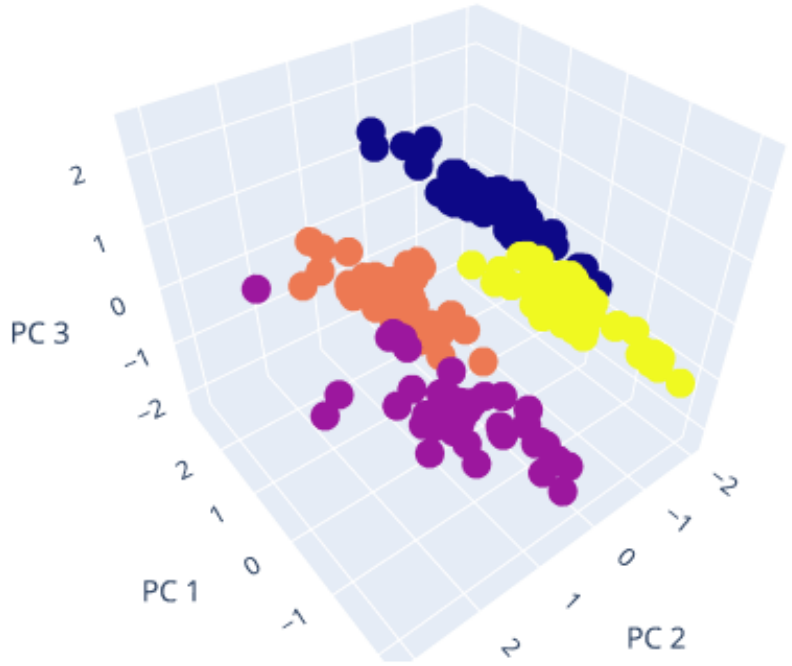

```
N_cluster: 2, score: 0.2518152915788437
N_cluster: 3, score: 0.26188419594665274
N_cluster: 4, score: 0.3012323168801352
N_cluster: 5, score: 0.3166442642857423
N_cluster: 6, score: 0.33622499051615934
N_cluster: 7, score: 0.35656464741976684
N_cluster: 8, score: 0.38733199737864654
N_cluster: 9, score: 0.40309165116555223
```



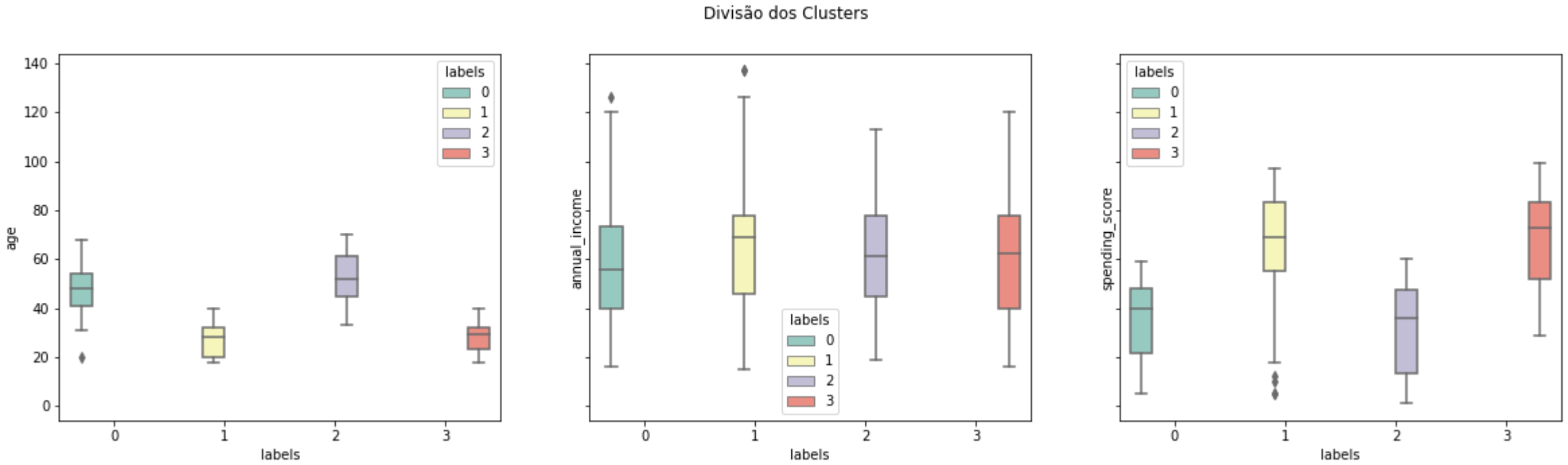
K-MEANS (Visualização com todas as entradas)



K-MEANS (Visualização 3D com todas as entradas)

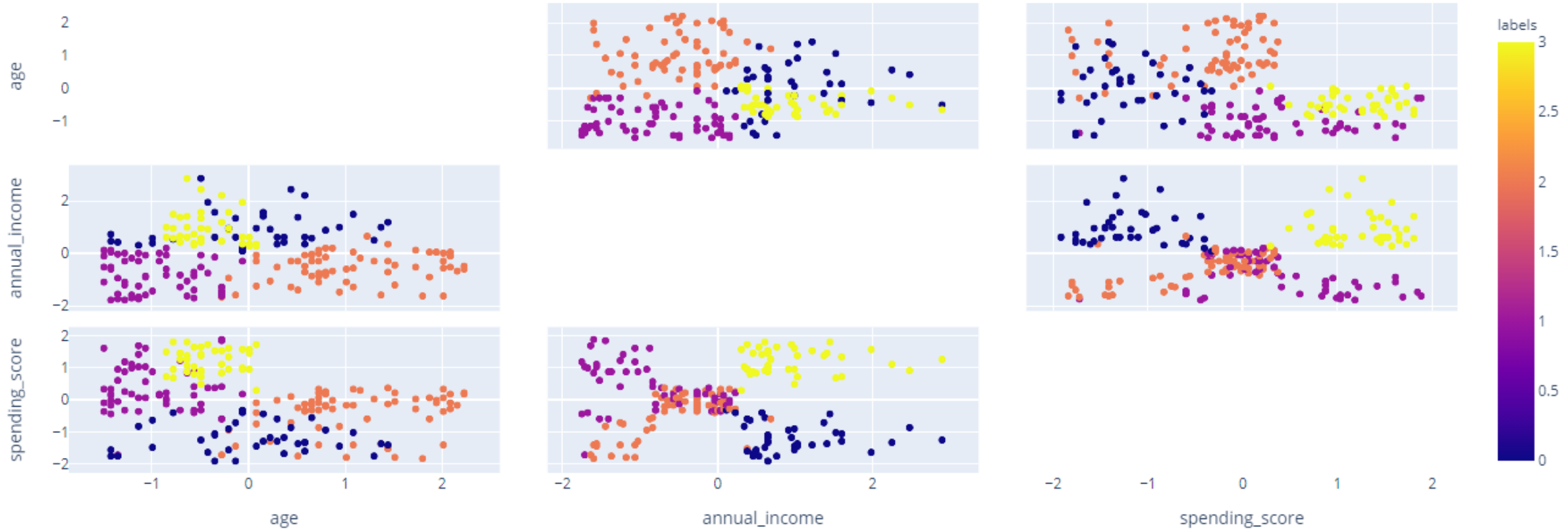


K-MEANS

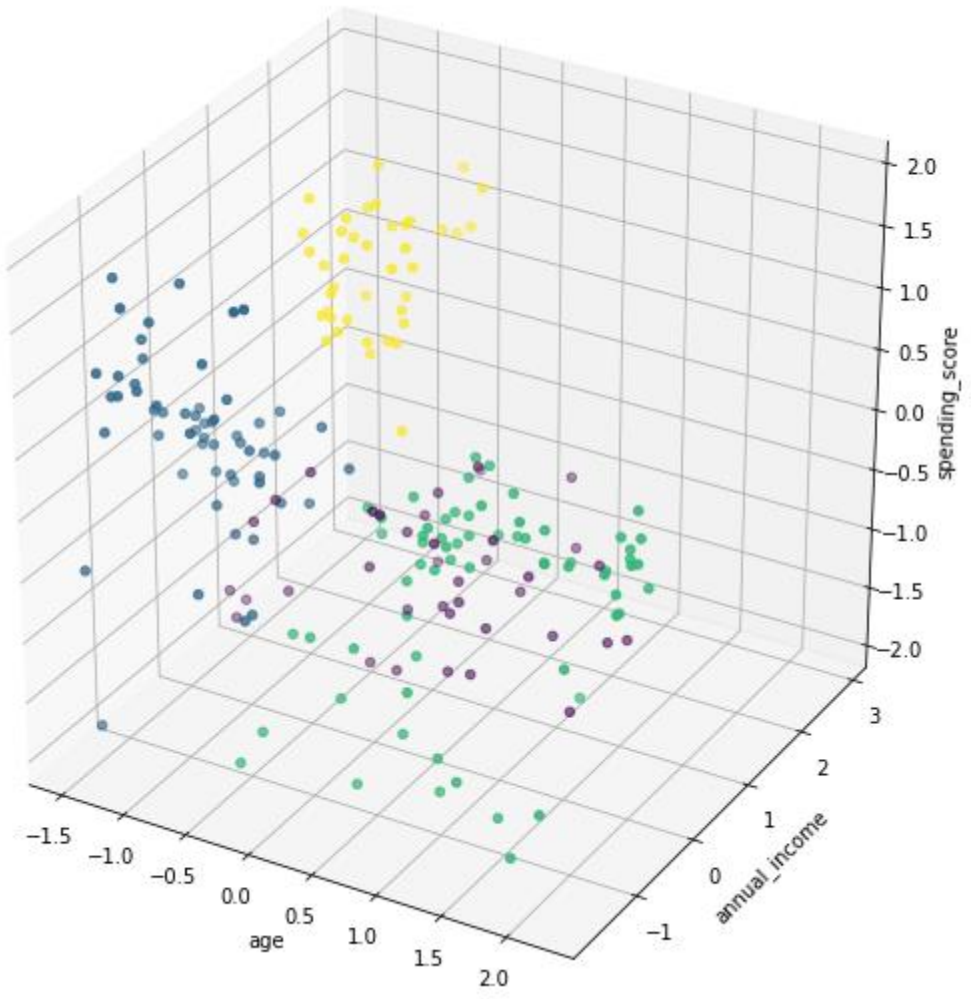


Clusters 0 e 3 compostos por pessoas dos gênero feminino e 1 e 2 do gênero masculino.

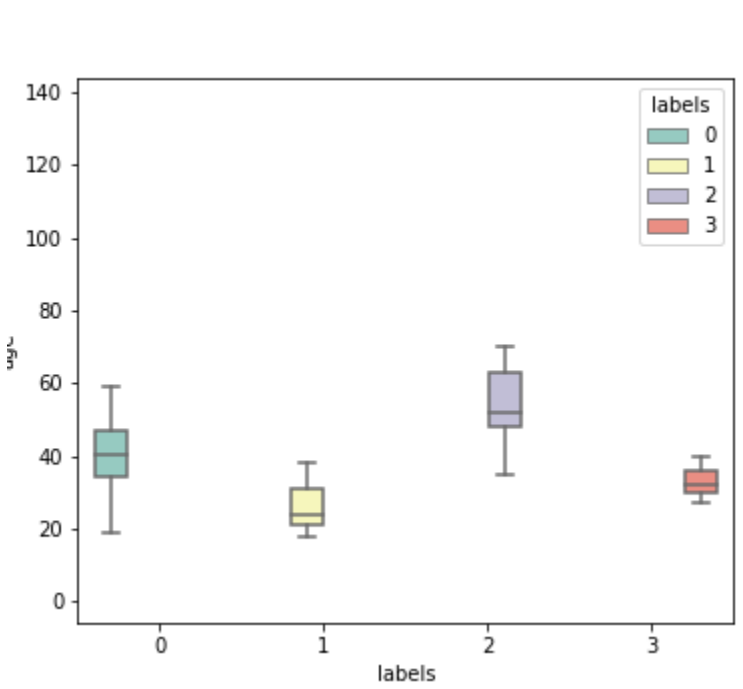
K-MEANS(Visualização de três entradas)



K-MEANS (Visualização 3D de três entradas)

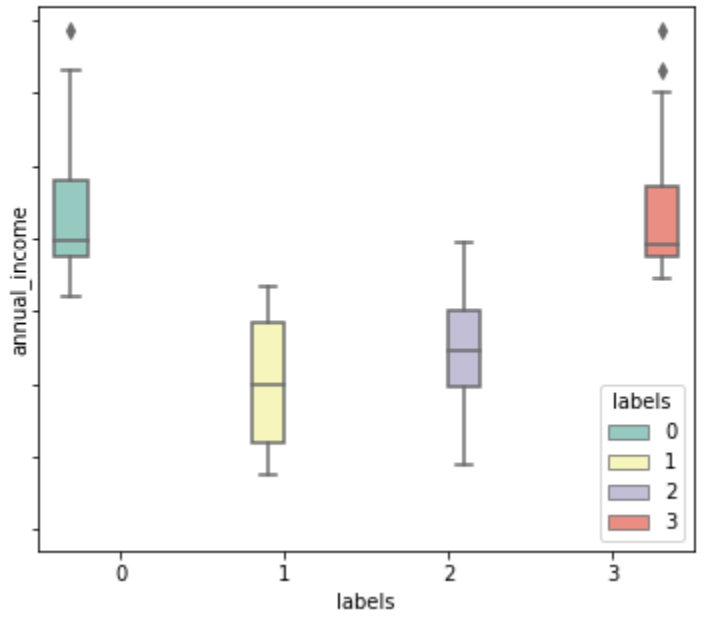


K-MEANS

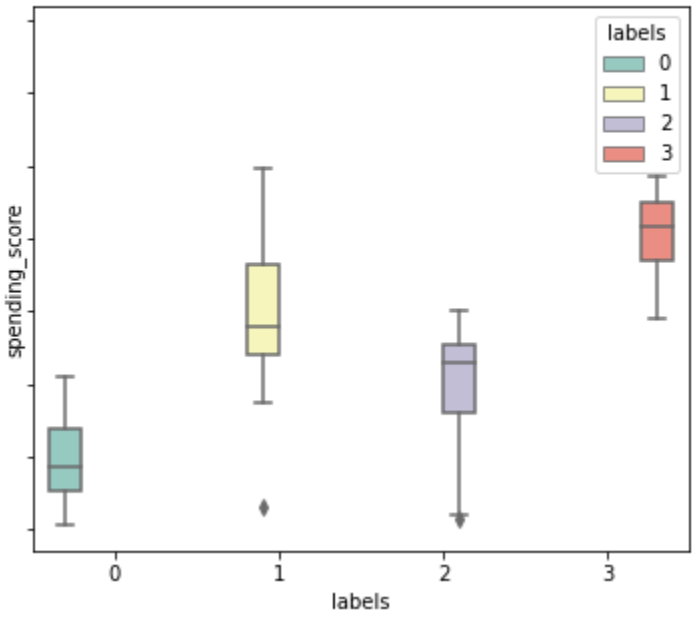


	count	mean	std	min	25%	50%	75%	max
labels								
0	38.0	39.37	10.62	19.0	34.0	40.5	46.75	59.0
1	57.0	25.44	5.71	18.0	21.0	24.0	31.00	38.0
2	65.0	53.98	9.42	35.0	48.0	52.0	63.00	70.0
3	40.0	32.88	3.86	27.0	30.0	32.0	36.00	40.0

Divisão dos Clusters



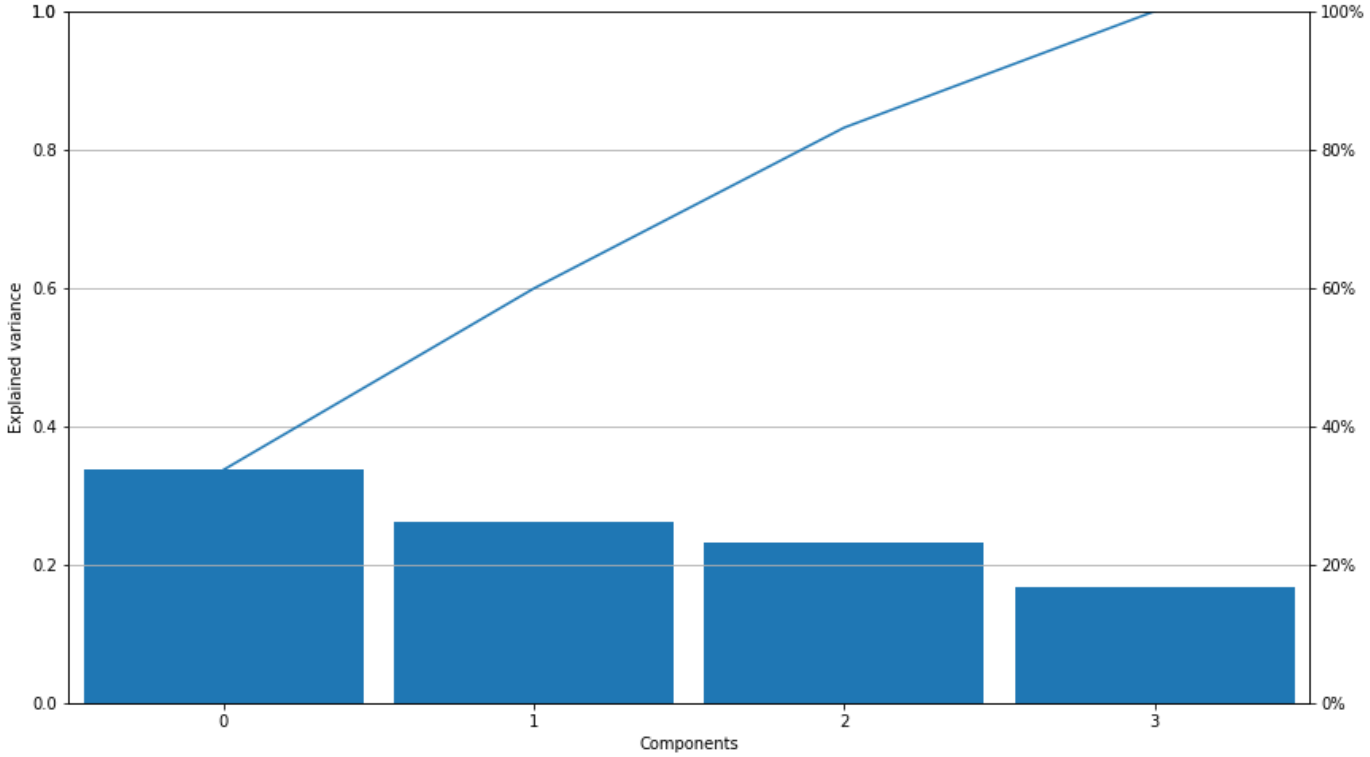
	count	mean	std	min	25%	50%	75%	max
labels								
0	38.0	86.50	16.76	64.0	75.25	79.5	96.0	137.0
1	57.0	40.00	17.03	15.0	24.00	40.0	57.0	67.0
2	65.0	47.71	14.65	18.0	39.00	49.0	60.0	79.0
3	40.0	86.10	16.34	69.0	74.75	78.5	94.0	137.0



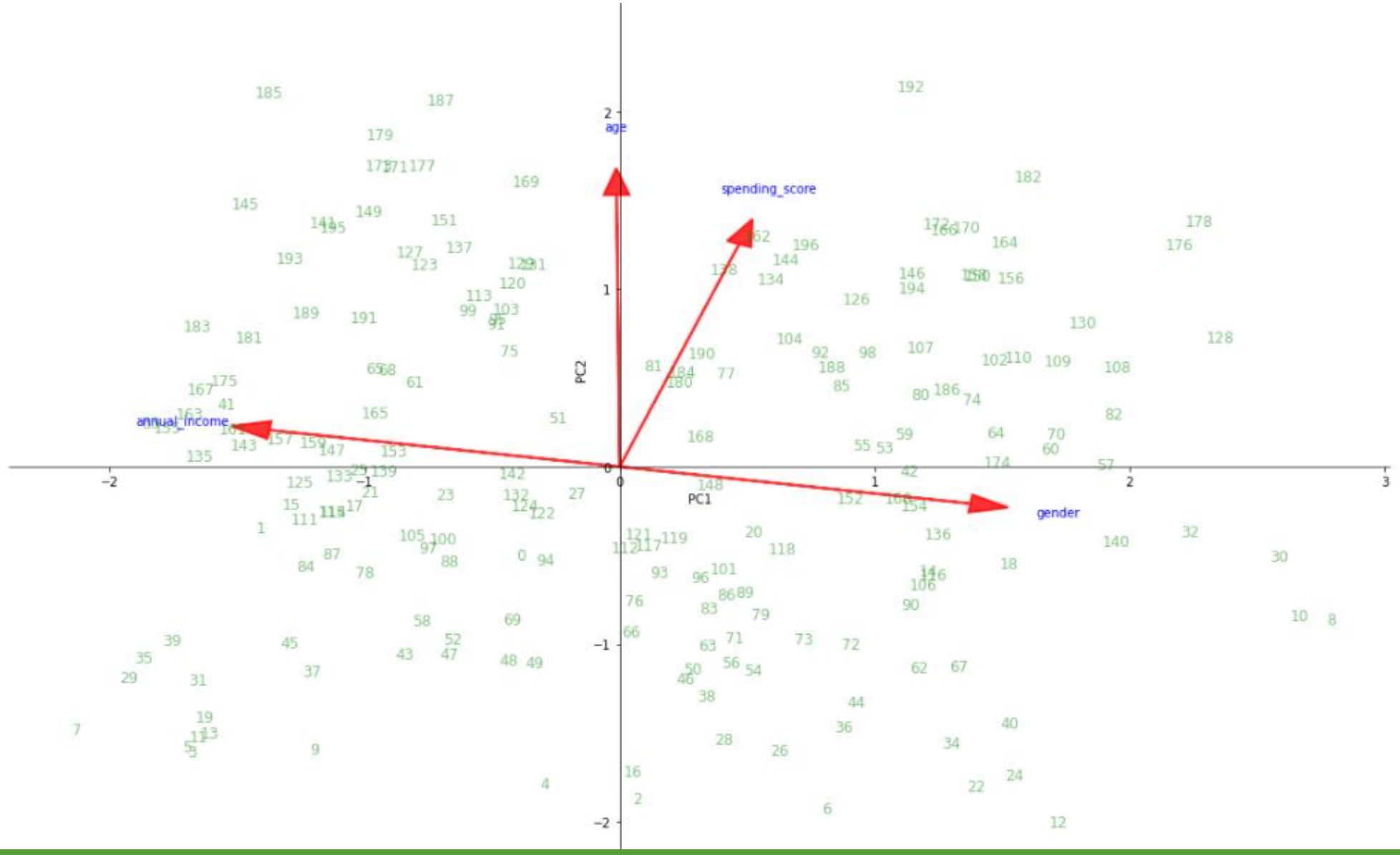
	count	mean	std	min	25%	50%	75%	max
labels								
0	38.0	19.58	11.68	1.0	10.5	17.0	27.75	42.0
1	57.0	60.30	18.43	6.0	48.0	56.0	73.00	99.0
2	65.0	39.97	16.41	3.0	32.0	46.0	51.00	60.0
3	40.0	81.53	10.00	58.0	74.0	83.0	90.00	97.0

PCA


age
annual_income
spending_score
gender
1.00
-0.09
-0.28
-0.09
1.00
-0.02
-0.28
-0.02
1.00
1.0
0.8
0.6
0.4
0.2
0.0
−0.2


Explained variance
Components
1.0
0.8
0.6
0.4
0.2
0.0
100%
80%
60%
40%
20%
0%
0
1
2
3

PCA (Componentes principais)



Resultados

Na primeira segmentação utilizando o algoritmo mais recomendado pela literatura, percebeu-se que independente do número de clusters selecionados a variável gênero sempre era a primeira variável a ser considerada, ou seja, o resultado foram dois clusters (0 e 3) compostos exclusivamente pelo gênero feminino e os outros dois (1 e 2) exclusivamente por gênero masculino. Além disso foi possível notar uma segmentação também por idade e por spending score. O cluster 0 e 3 possuem idades médias e baixas e os 1 e 2 possuem idades médias e altas. Quanto ao spending score o cluster 0 e 2 apresentam médio e baixo valor nessa variável e o 3 apresenta valores médios e altos, o cluster 1 também apresenta valores médios e altos em sua maioria, no entanto possui alguns outliers de valores baixos.

O algoritmo do K-means faz a clusterização sempre pela distância euclidiana, e com isso, tendo varáveis categóricas, como gênero, ele acabou enviesando o processo fazendo com que a clusterização sempre levasse consideração o ponto máximo e mínimo da variável categórica. Por isso fizemos uma nova aplicação do K-means retirando a variável categórica. Como resultado tivemos clusters com ambos os gêneros. Comparado ao dataset original os clusters 1 e 2 tiveram uma proporção levemente maior de gênero femino, enquanto o 0 e principalmente o 3 apresentaram uma relação maior de gênero masculino. Todos apresentaram baixa dispersão na variável idade.

No spending score o cluster 3 possui o maior nível de gastos, acompanhado pelo cluster 1. O Cluster tem um spending score intermediário entre os maiores valores e os menores do cluster 0. Se tratando de renda anual os clusters 0 e 3 possuem comportamento muito semelhante em todos os aspectos, possuindo os maiores valores, já os clusters 1 e 2 possuem valores médios e intermediários, com o cluster 1 apresentando uma maior variabilidade. Quanto a idade o cluster 2 possui a maior média (53,98), o valor mínimo de idade nesse cluster é próximo da idade máxima dos clusters 1 e 3.

Foi analisado a possibilidade de reduzir as dimensões utilizando o PCA, no entanto os resultados não foram positivos. Trazendo um grande aumento na perda de informação em um dataset que já possuía inicialmente um volume limitado.

CONCLUSÃO

Um dos métodos convencionais de agrupamento comumente usados em técnicas de agrupamento e eficientemente usados para grandes dados é o algoritmo K-Means. No entanto, verificamos que seu método não é o mais adequado para dados que contenham variáveis categóricas. Esse problema ocorre quando a função de custo em K-Médias é calculada usando a distância euclidiana que só é adequada para dados numéricos.

Diante desses problemas, Huang propôs um algoritmo denominado K-Prototype, com o objetivo de evitar o desvio do valor do resultado do agrupamento do Recurso Categórico ou do Recurso Numérico e controlar o peso relativo da dissimilaridade entre eles.

Em análises realizadas futuras aconselhamos a utilização do K-prototype para conseguir fazer uma análise mais precisa dos dados com variável categórica.

OBRIGADO!