CÁC GIẢI THUẬT PHÂN CỤM CƠ BẢN

3.1. GIẢI THUẬT K-MEANS

3.1.1. Ôn tập lý thuyết

3.1.2. Bài tập mẫu

3.1.3. Bài tập thực hành 1

Xây dựng mô hình phân cụm K-means trên tập dữ liệu chim cánh cụt. Dữ liệu lấy tại https://www.kaggle.com/code/youssefaboelwafa/clustering-penguins-species-k-means-clustering

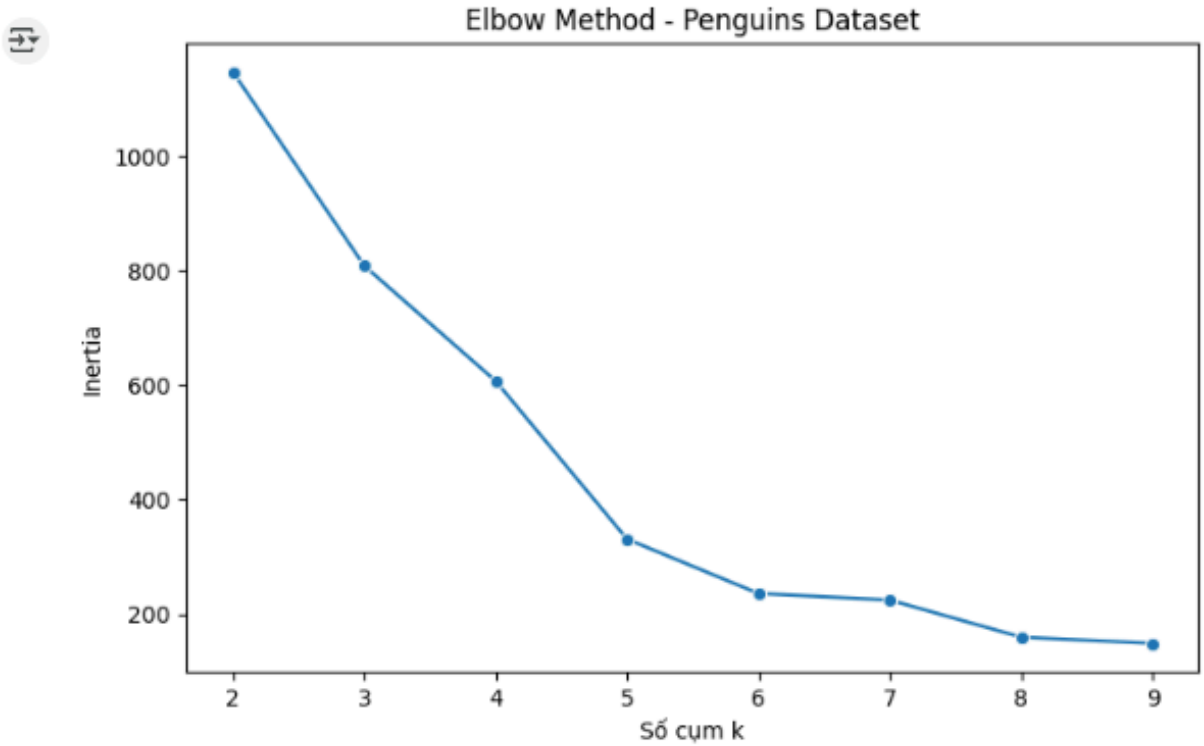
1. Import thư viện
2. Đọc dữ liệu
3. Tiền xử lý dữ liệu

```
Kích thước dữ liệu: (344, 5)
   culmen_length_mm  culmen_depth_mm  flipper_length_mm  body_mass_g     sex
0              39.1             18.7              181.0       3750.0    MALE
1              39.5             17.4              186.0       3800.0  FEMALE
2              40.3             18.0              195.0       3250.0  FEMALE
3               NaN              NaN                NaN          NaN     NaN
4              36.7             19.3              193.0       3450.0  FEMALE
culmen_length_mm     2
culmen_depth_mm      2
flipper_length_mm    2
body_mass_g          2
sex                  9
dtype: int64
<class 'pandas.core.frame.DataFrame'>
Index: 335 entries, 0 to 343
Data columns (total 5 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   culmen_length_mm   335 non-null    float64
 1   culmen_depth_mm    335 non-null    float64
 2   flipper_length_mm  335 non-null    float64
 3   body_mass_g        335 non-null    float64
 4   sex                335 non-null    int64
dtypes: float64(4), int64(1)
```

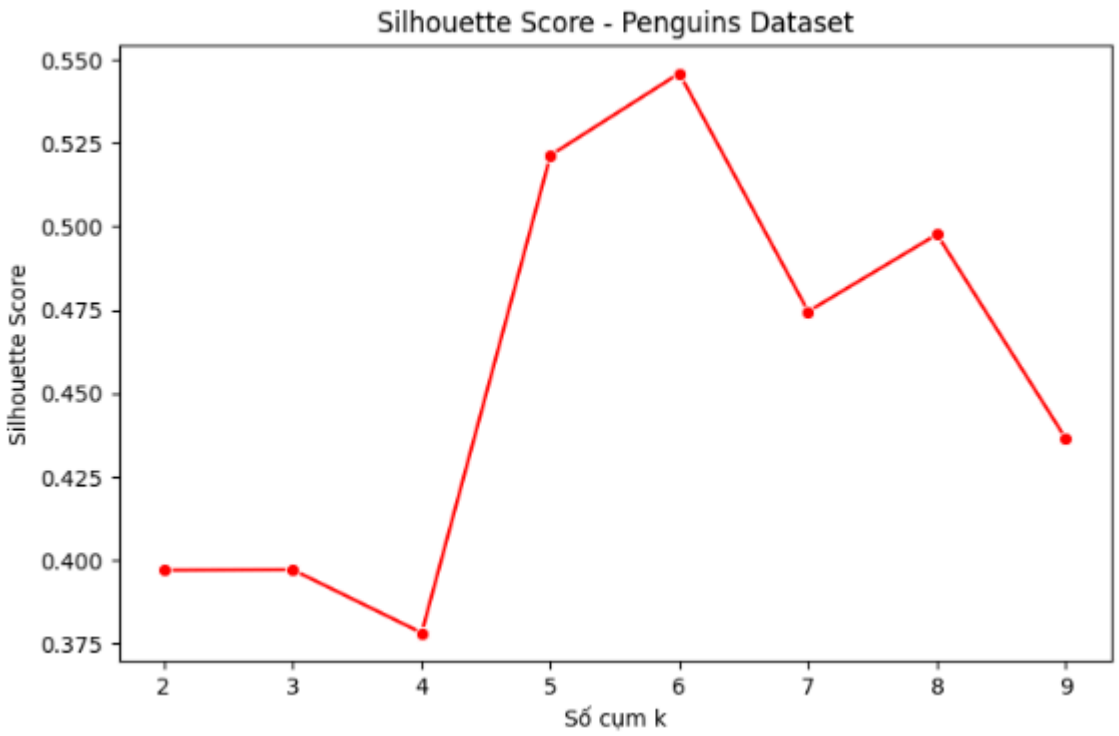
Xóa hàng thiếu dữ liệu
Encode biến phân loại (species, island, sex)

```
---  ------             --------------  -----
 0   culmen_length_mm   335 non-null    float64
 1   culmen_depth_mm    335 non-null    float64
 2   flipper_length_mm  335 non-null    float64
 3   body_mass_g        335 non-null    float64
 4   sex                335 non-null    int64
dtypes: float64(4), int64(1)
memory usage: 15.7 KB
Sau khi xử lý: None
k = 2, Inertia = 1145.96, Silhouette = 0.397
k = 3, Inertia = 807.61, Silhouette = 0.397
k = 4, Inertia = 606.50, Silhouette = 0.378
k = 5, Inertia = 330.57, Silhouette = 0.521
k = 6, Inertia = 235.69, Silhouette = 0.546
k = 7, Inertia = 224.19, Silhouette = 0.475
k = 8, Inertia = 159.44, Silhouette = 0.498
k = 9, Inertia = 148.63, Silhouette = 0.436
```

4. Xây dựng mô hình KMeans với nhiều k
5. Chuẩn hóa dữ liệu
6. Vẽ Elbow Method



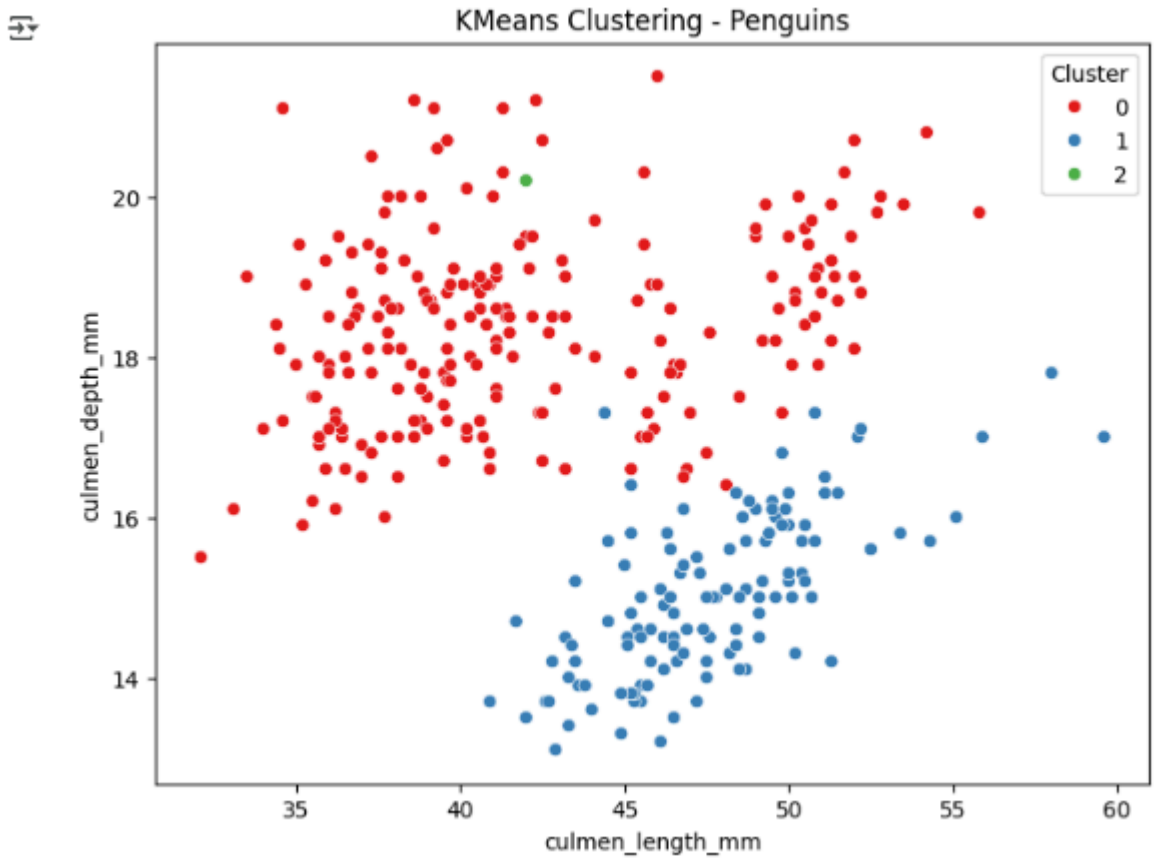
7. Vẽ Silhouette Score



8. Phân tích cụm

Cluster	culmen_length_mm	culmen_depth_mm	flipper_length_mm	body_mass_g	sex
0	41.929577	18.373239	190.422535	3714.788732	1.502347
1	47.628926	15.025620	216.933884	5079.132231	1.495868
2	42.000000	20.200000	5000.000000	4250.000000	2.000000

9. Trực quan hóa kết quả theo 2 đặc trưng chính

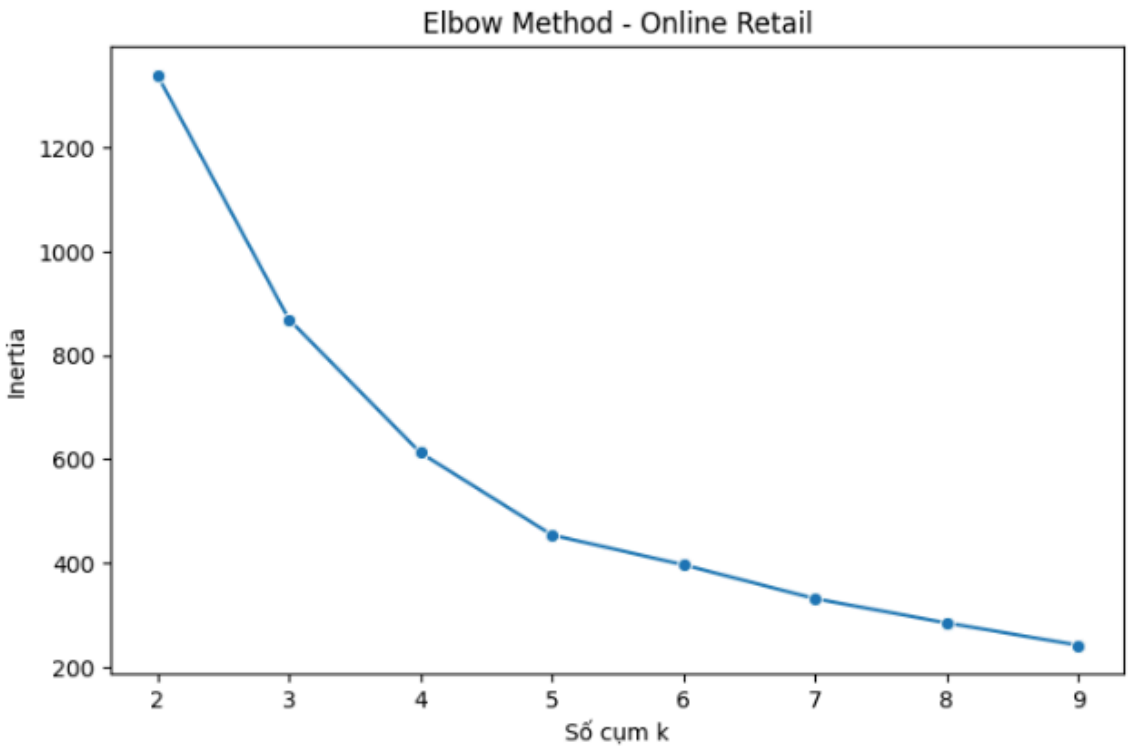


3.1.4. Bài tập thực hành 2

Xây dựng mô hình phân cụm K-means trên tập dữ liệu mua sắm tại siêu thị. Dữ liệu lấy tại https://www.kaggle.com/datasets/hellbuoy/online-retail-customer-clustering

```
Kích thước dữ liệu: (24721, 8)
  InvoiceNo StockCode                          Description  Quantity  \
0    536365    85123A   WHITE HANGING HEART T-LIGHT HOLDER         6
1    536365     71053                  WHITE METAL LANTERN         6
2    536365    84406B       CREAM CUPID HEARTS COAT HANGER         8
3    536365    84029G  KNITTED UNION FLAG HOT WATER BOTTLE         6
4    536365    84029E       RED WOOLLY HOTTIE WHITE HEART.         6

         InvoiceDate  UnitPrice  CustomerID         Country
0  01-12-2010 08:26       2.55     17850.0  United Kingdom
1  01-12-2010 08:26       3.39     17850.0  United Kingdom
2  01-12-2010 08:26       2.75     17850.0  United Kingdom
3  01-12-2010 08:26       3.39     17850.0  United Kingdom
4  01-12-2010 08:26       3.39     17850.0  United Kingdom

RFM head:
   CustomerID  Recency  Frequency  Monetary
0     12347.0       92         31    711.79
1     12386.0       62          8    258.90
2     12395.0      214         12    346.10
3     12427.0      215         10    303.50
4     12429.0       31         20   1281.50


RFM head:
   CustomerID  Recency  Frequency  Monetary
0     12347.0       92         31    711.79
1     12386.0       62          8    258.90
2     12395.0      214         12    346.10
3     12427.0      215         10    303.50
4     12429.0       31         20   1281.50
k=2: Inertia=1339.68, Silhouette=0.432
k=3: Inertia=869.14, Silhouette=0.466
k=4: Inertia=612.12, Silhouette=0.469
k=5: Inertia=453.76, Silhouette=0.484
k=6: Inertia=396.32, Silhouette=0.480
k=7: Inertia=331.12, Silhouette=0.408
k=8: Inertia=284.45, Silhouette=0.415
k=9: Inertia=241.56, Silhouette=0.431
```

[↓]


Elbow Method - Online Retail
Inertia
Số cụm k
1200
1000
800
600
400
200
2
3
4
5
6
7
8
9

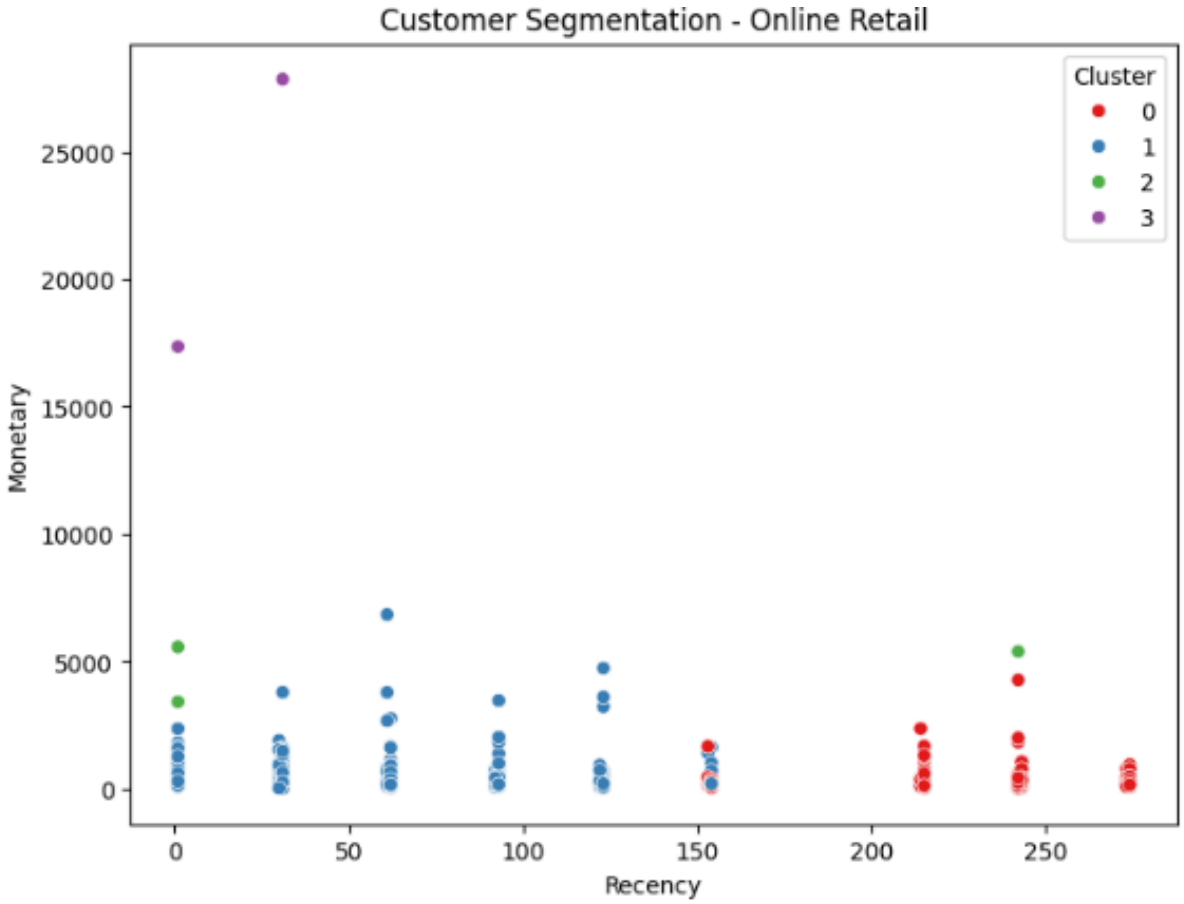
[↓]


Silhouette Score - Online Retail
Silhouette Score
Số cụm k
0.48
0.47
0.46
0.45
0.44
0.43
0.42
0.41
2
3
4
5
6
7
8
9

Đặc trưng trung bình của từng cụm:

Cluster	CustomerID	Recency	Frequency	Monetary
0	15477.114155	233.242009	18.255708	373.404566
1	15511.419271	70.992188	27.377604	512.261979
2	15169.666667	81.333333	351.000000	4792.390000
3	16581.500000	16.000000	63.500000	22589.695000

[+]



3.2. GIẢI THUẬT PHÂN CỤM ĐA CẤP

3.2.1. Ôn tập lý thuyết

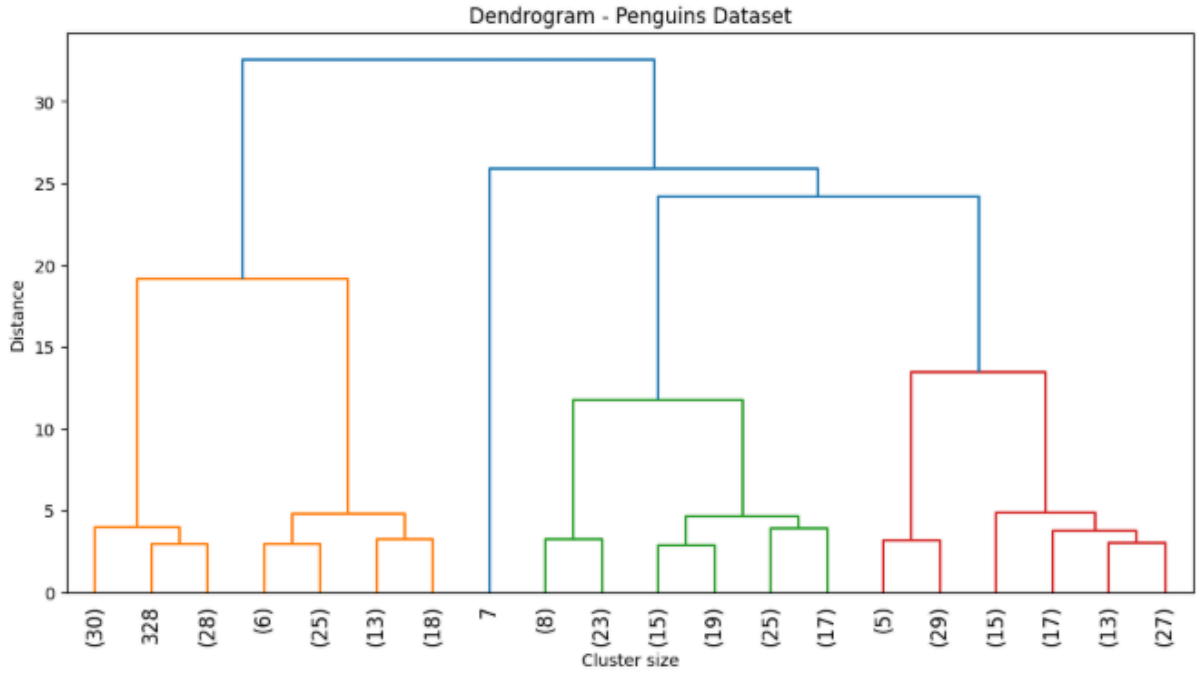
3.2.2. Bài làm mẫu

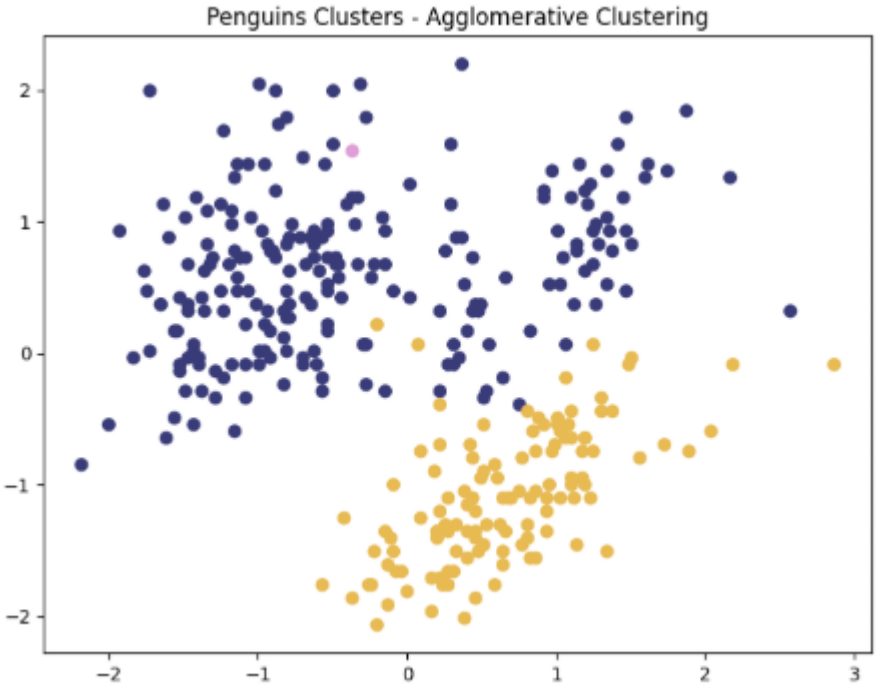
3.2.3. Bài tập thực hành 1

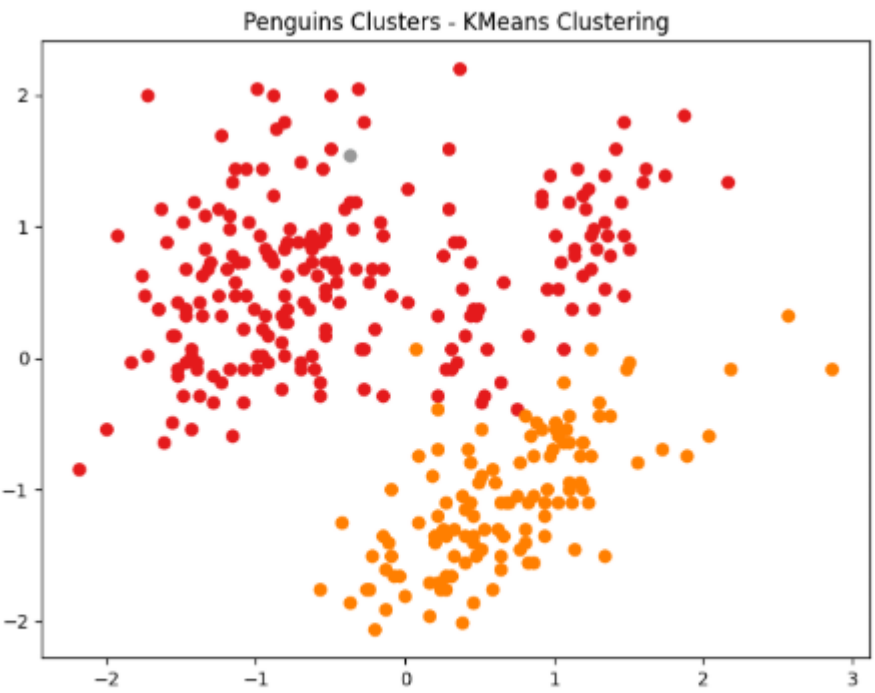
Xây dựng mô hình phân cụm đa cấp trên tập dữ liệu chim cánh cụt. Dữ liệu lấy tại
https://www.kaggle.com/code/youssefaboelwafa/clustering-penguins-species-k-means-clustering

```
Kích thước dữ liệu: (344, 5)
   culmen_length_mm  culmen_depth_mm  flipper_length_mm  body_mass_g     sex
0              39.1             18.7              181.0       3750.0    MALE
1              39.5             17.4              186.0       3800.0  FEMALE
2              40.3             18.0              195.0       3250.0  FEMALE
3               NaN              NaN                NaN          NaN     NaN
4              36.7             19.3              193.0       3450.0  FEMALE
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 344 entries, 0 to 343
Data columns (total 5 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   culmen_length_mm   342 non-null    float64
 1   culmen_depth_mm    342 non-null    float64
 2   flipper_length_mm  342 non-null    float64
 3   body_mass_g        342 non-null    float64
 4   sex                335 non-null    object
dtypes: float64(4), object(1)
memory usage: 13.6+ KB
None

Số lượng giá trị thiếu:
 culmen_length_mm     2
culmen_depth_mm      2
flipper_length_mm    2
body_mass_g          2
sex                  9
dtype: int64

Sau khi mã hóa & làm sạch: (335, 5)
```

Sau khi mã hóa & làm sạch: (335, 5)





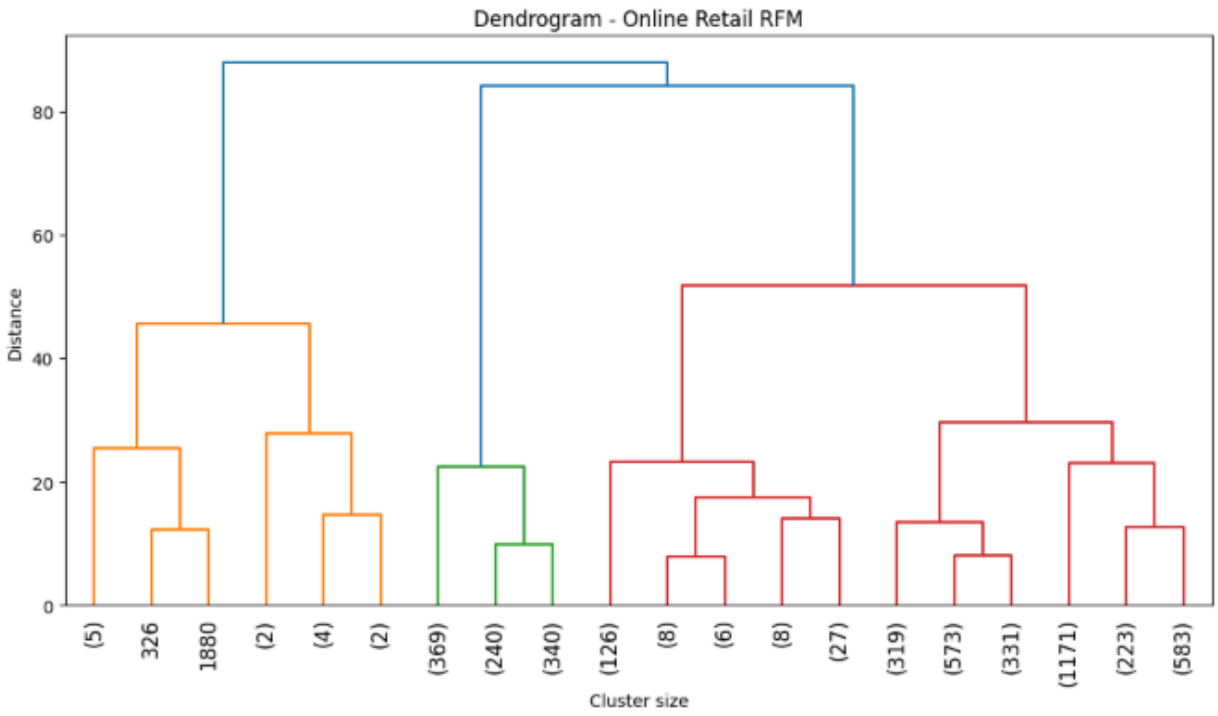


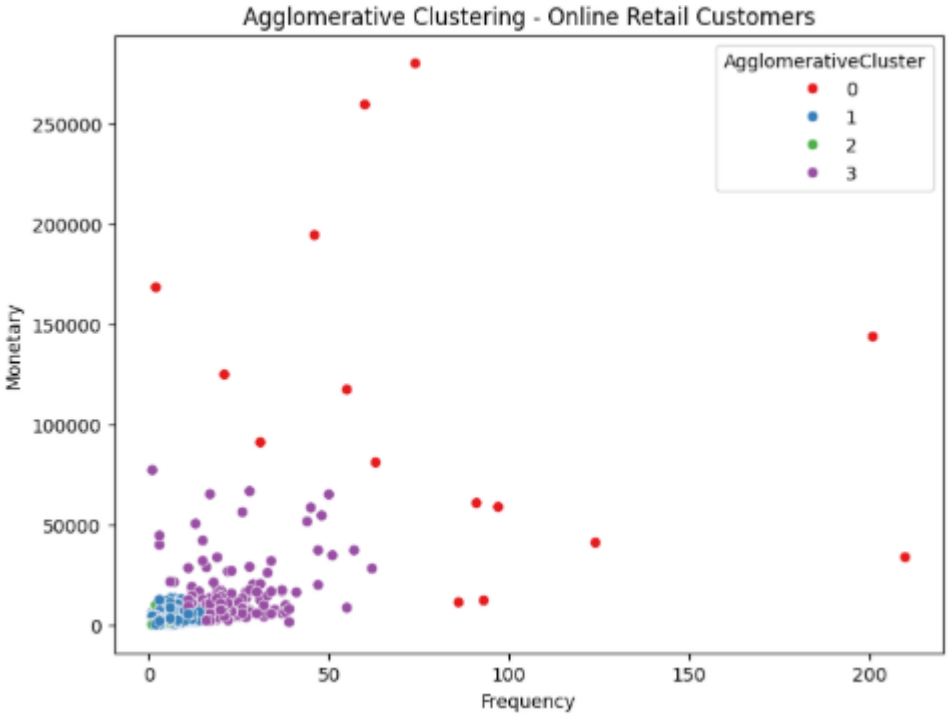

```
Silhouette Scores:
Agglomerative Clustering: 0.39560631008961056
KMeans Clustering: 0.397044474123945
```

3.2.4. Bài tập thực hành 2

Xây dựng mô hình phân cụm đa cấp trên tập dữ liệu mua sắm tại siêu thị. Dữ liệu lấy tại https://www.kaggle.com/datasets/hellbuoy/online-retail-customer-clustering

```
Kích thước dữ liệu: (541909, 8)
  InvoiceNo StockCode                          Description  Quantity  \
0    536365    85123A   WHITE HANGING HEART T-LIGHT HOLDER         6
1    536365     71053                  WHITE METAL LANTERN         6
2    536365    84406B       CREAM CUPID HEARTS COAT HANGER         8
3    536365    84029G  KNITTED UNION FLAG HOT WATER BOTTLE         6
4    536365    84029E       RED WOOLLY HOTTIE WHITE HEART.         6

        InvoiceDate  UnitPrice  CustomerID         Country
0  01-12-2010 08:26       2.55     17850.0  United Kingdom
1  01-12-2010 08:26       3.39     17850.0  United Kingdom
2  01-12-2010 08:26       2.75     17850.0  United Kingdom
3  01-12-2010 08:26       3.39     17850.0  United Kingdom
4  01-12-2010 08:26       3.39     17850.0  United Kingdom

--- RFM sample ---
   CustomerID  Recency  Frequency  Monetary
0     12346.0      326          1  77183.60
1     12347.0        2          7   4310.00
2     12348.0       75          4   1797.24
3     12349.0       19          1   1757.55
4     12350.0      310          1    334.40
```






```
Silhouette Scores:
Agglomerative Clustering: 0.615110241323992
KMeans Clustering: 0.6161144819517276
```