**In-Class Exercise: Reading Clustering Output**

The following is the output from a clustering analysis in SAS Enterprise Miner. It uses a customer information database from a bank to analyze distinct customer groups.

Five pieces of customer data were used to create the clusters:

CRSCORE – The customer’s credit score.
ATMAMT – The ATM withdrawal amount.
INCOME – The customer’s annual income.
HMVAL – The value of the customer’s home.
AGE – The customer’s age.

Here is the result of the clustering analysis. We first asked SAS to group the data into only three clusters:

**Scenario #1: Three Clusters**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Segment (Cluster) ID** | **Frequency (size) of cluster** | **Root Mean Squared Standard Deviation** | **Distance to Nearest Cluster** | **AGE** | **ATMAMT** | **CRSCORE** | **HMVAL** | **INCOME** |
| 1 | 12167 | 0.899116 | 11.99024 | 47.94117 | 1048.115 | 663.8654 | 110.8019 | 40.89706 |
| 2 | 699 | 2.390794 | 11.99024 | 59.94118 | 1041.21 | 713.2565 | 380.6111 | 63.16667 |
| 3 | 41 | 3.013538 | 12.61584 | 50.91176 | 47162.73 | 676.6585 | 116.1471 | 36.52941 |

Answer the questions about this output:

1. How many distinct customer groups (segments) are there?
2. Explain how the customers in cluster 1 are different from cluster 2?
3. What aspect of the customer data most differentiates cluster 1 from cluster 3?
4. Which cluster has the highest cohesion? In practical terms, what does that mean?

**<<NEXT PAGE>>**

We asked SAS to take the same data and split the customers into 10 groups instead of three.

**Scenario #2: Ten Clusters**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Segment (Cluster) ID** | **Frequency (size) of cluster** | **Root Mean Squared Standard Deviation** | **Distance to Nearest Cluster** | **AGE** | **ATMAMT** | **CRSCORE** | **HMVAL** | **INCOME** |
| 1 | 678 | 0.777343 | 1.85956 | 58.82796 | 1633.386 | 686.0015 | 145.6862 | 63.20502 |
| 2 | 2027 | 0.735691 | 1.85956 | 39.45261 | 809.9025 | 653.7394 | 135.0763 | 86.90867 |
| 3 | 409 | 1.530902 | 10.26644 | 53.4 | 130.4035 | 673.1138 | 535.8 | 70.4 |
| 4 | 36 | 1.34275 | 6.257339 | 52.25 | 34061.55 | 684.7778 | 120.1071 | 33.78571 |
| 5 | 4552 | 0.569064 | 2.052375 | 39.26974 | 712.3268 | 639.399 | 104.8276 | 30.76862 |
| 6 | 313 | 0.961229 | 3.098761 | 48.26953 | 11286.11 | 668.1661 | 114.6107 | 41.84733 |
| 7 | 4 | 2.590545 | 8.522712 | 38 | 100937 | 684 | 110.6667 | 40.33333 |
| 8 | 313 | 1.495353 | 7.324831 | 57.6 | 1028.864 | 605.8506 | 305.75 | 70.3125 |
| 9 | 4569 | 0.590474 | 2.052375 | 59.24229 | 758.1439 | 699.227 | 101.8305 | 28.5347 |
| 10 | 6 | 1.59657 | 8.522712 | 51.33333 | 70070.36 | 654.8333 | 110 | 44.33333 |

Now answer the following questions:

1. Is the root mean squared standard deviation of these clusters higher or lower than they were in the three cluster scenario? Why?
2. Is the distance to the nearest cluster higher or lower than in the three cluster scenario? Why?
3. Which scenario (#1 or #2) has higher cohesion among its clusters?
4. Which scenario (#1 or #2) has higher separation between its clusters?