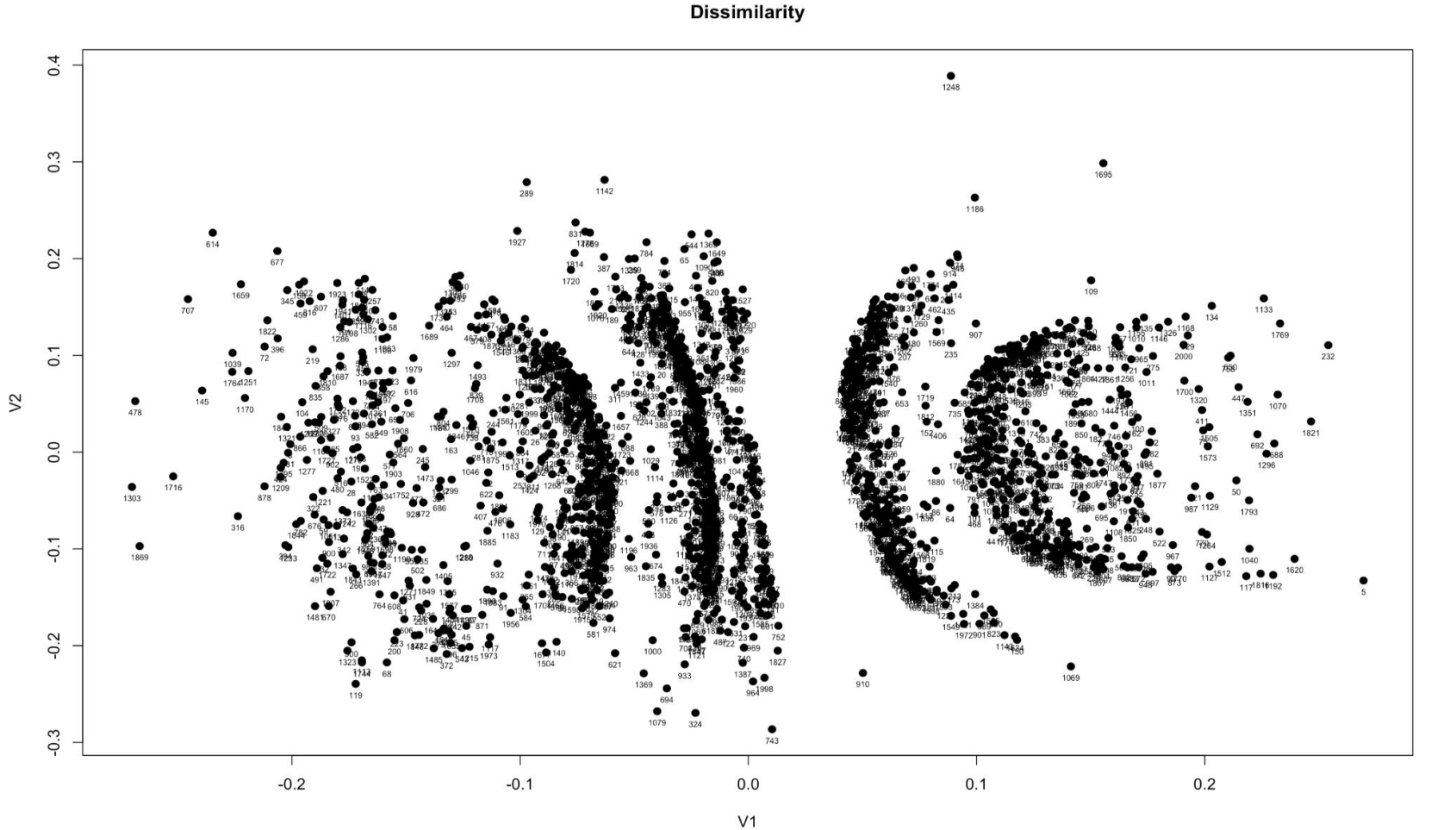
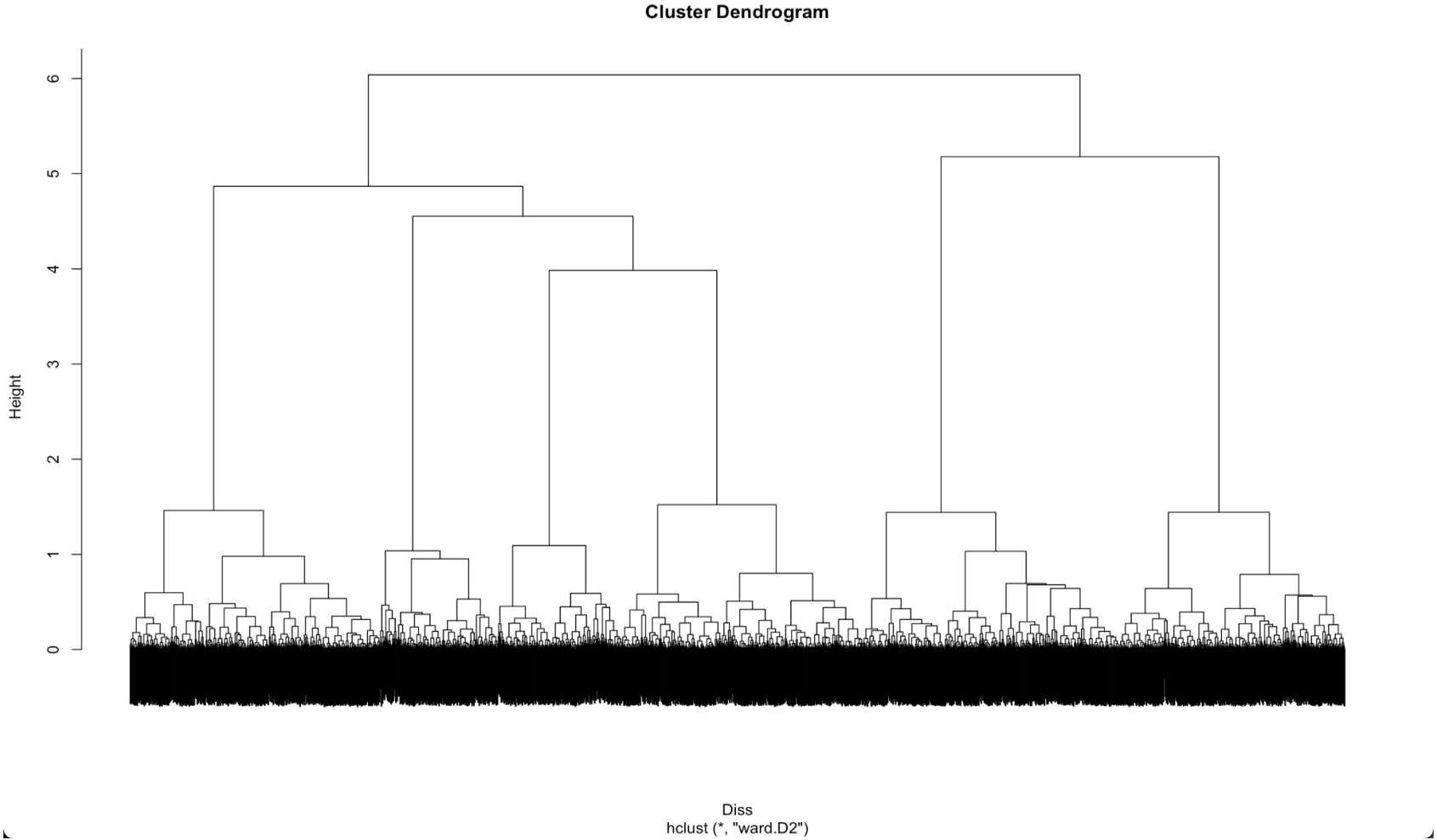
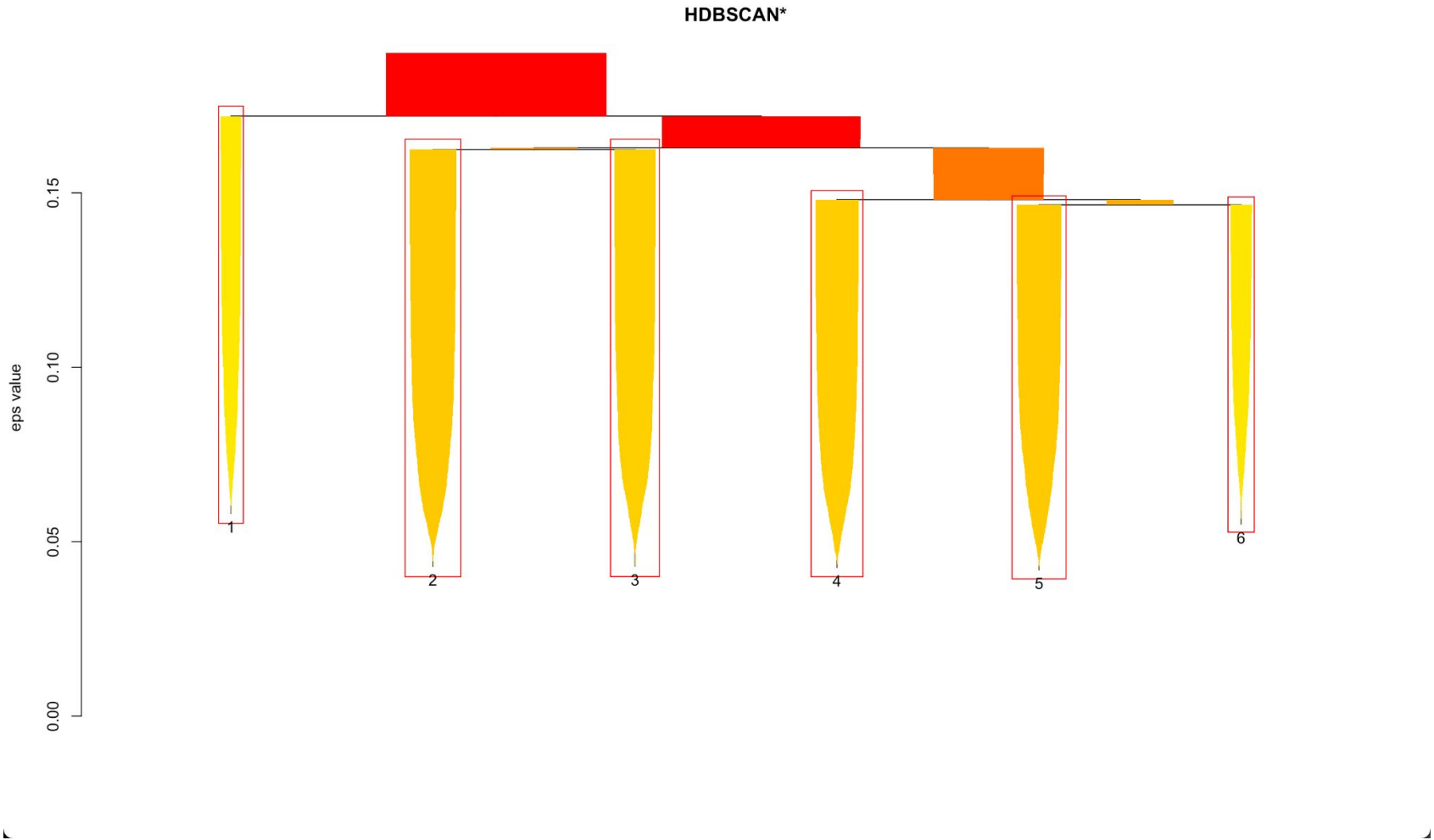
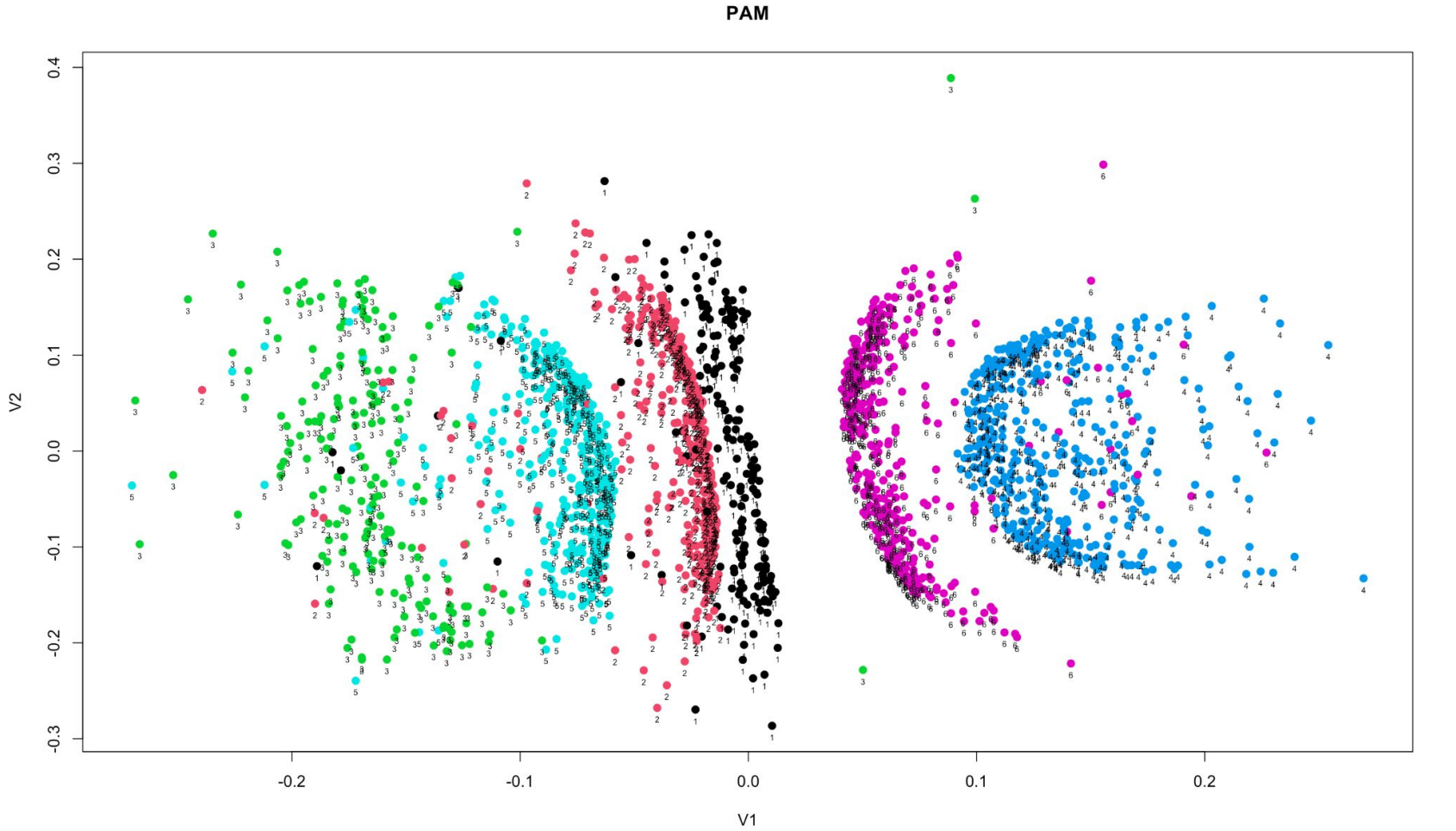
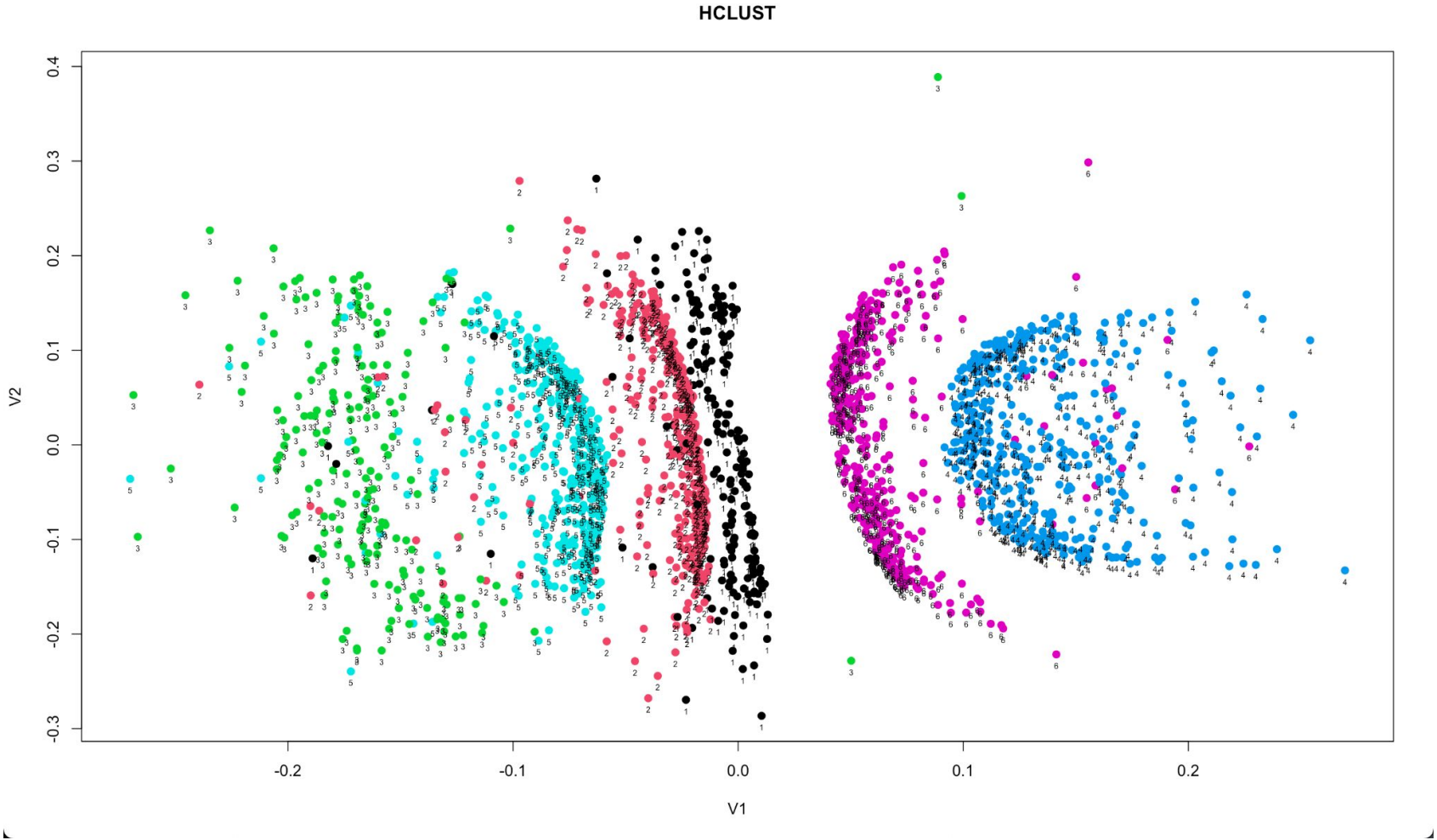
Predict Earnings

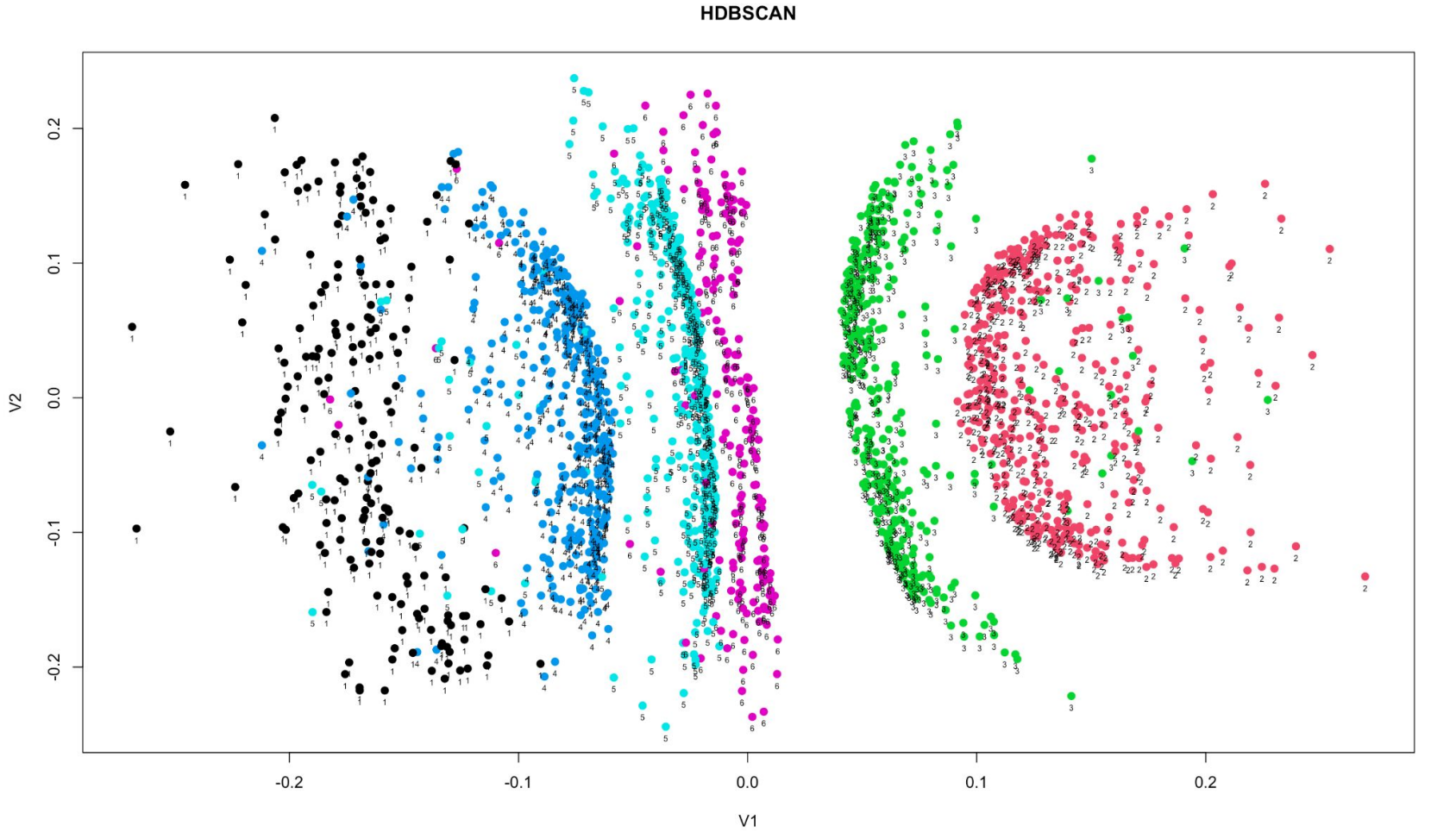

Dissimilarity
V2
V1
0.4
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
-0.2
-0.1
0.0
0.1
0.2

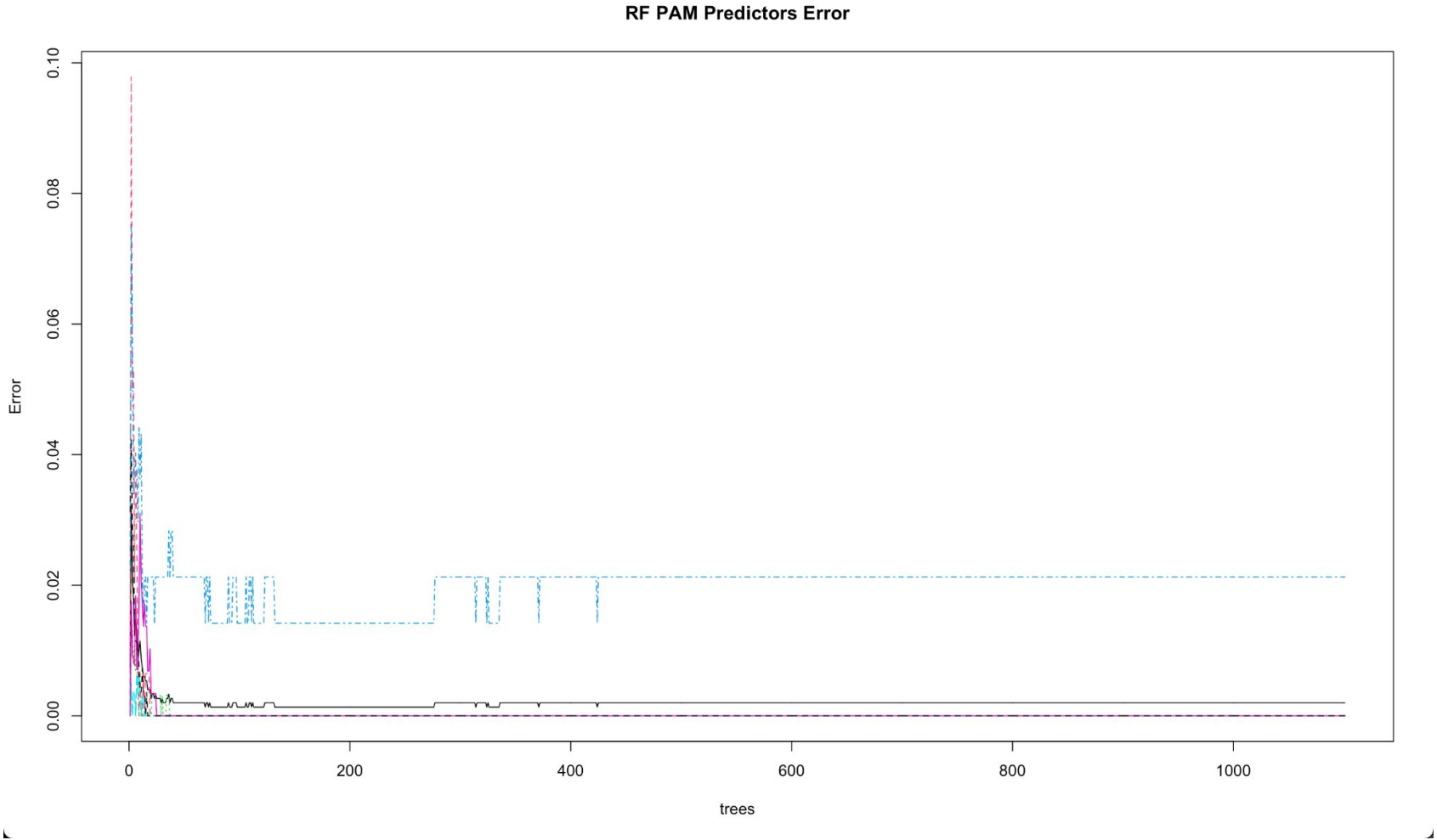

Cluster Dendrogram
Height
0
1
2
3
4
5
6
Diss
hclust (*, "ward.D2")

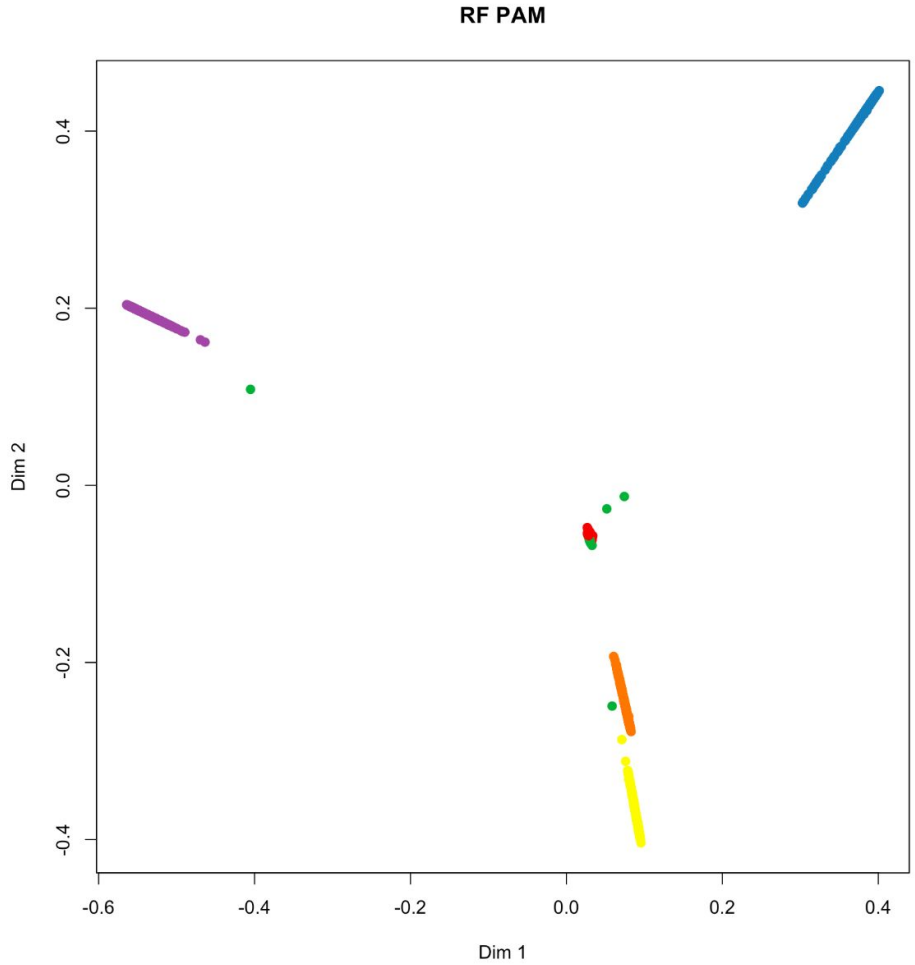

HDBSCAN*
eps value
0.15
0.10
0.05
0.00
1
2
3
4
5
6

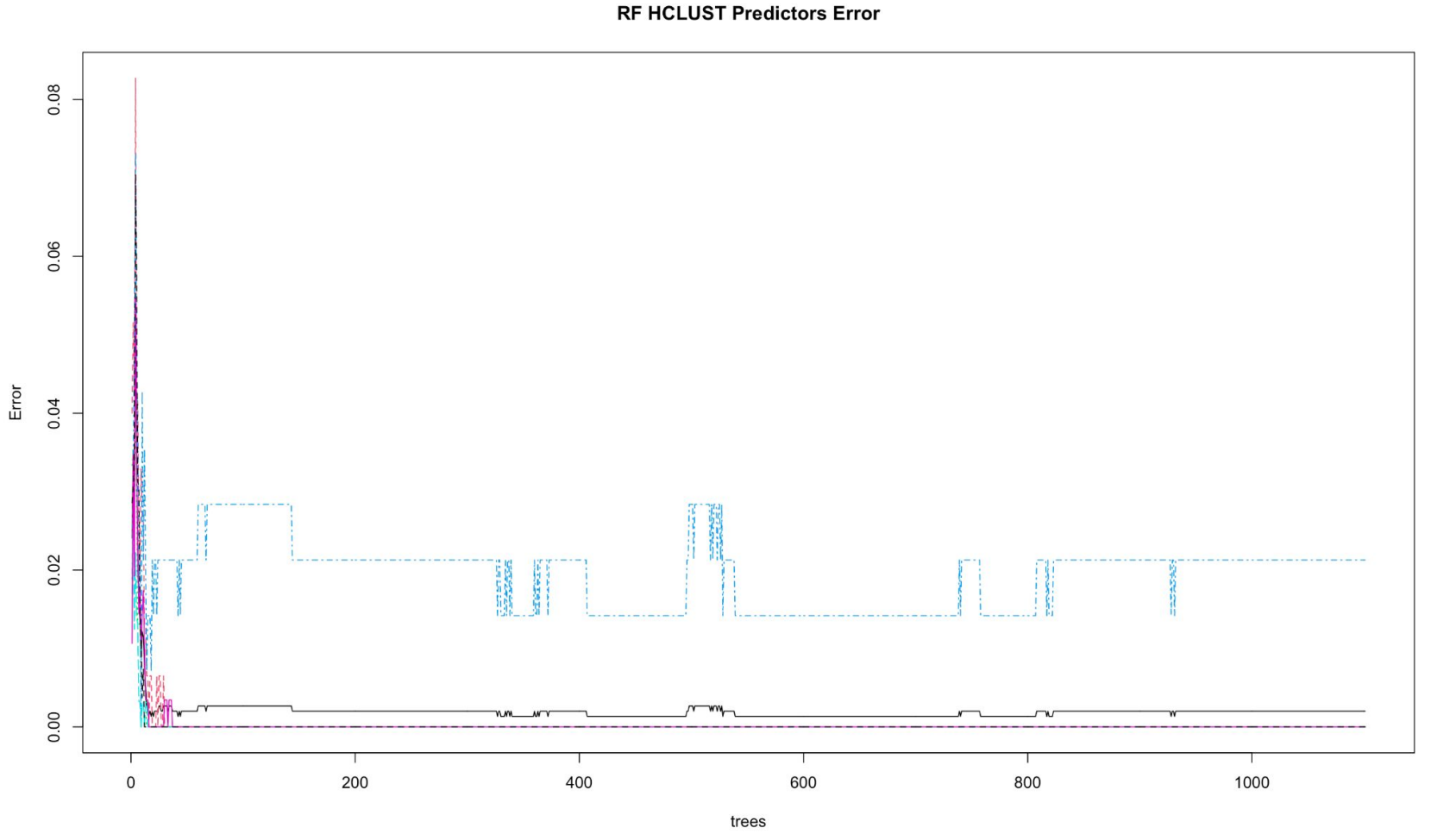

PAM
V2
V1
0.4
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
-0.2
-0.1
0.0
0.1
0.2

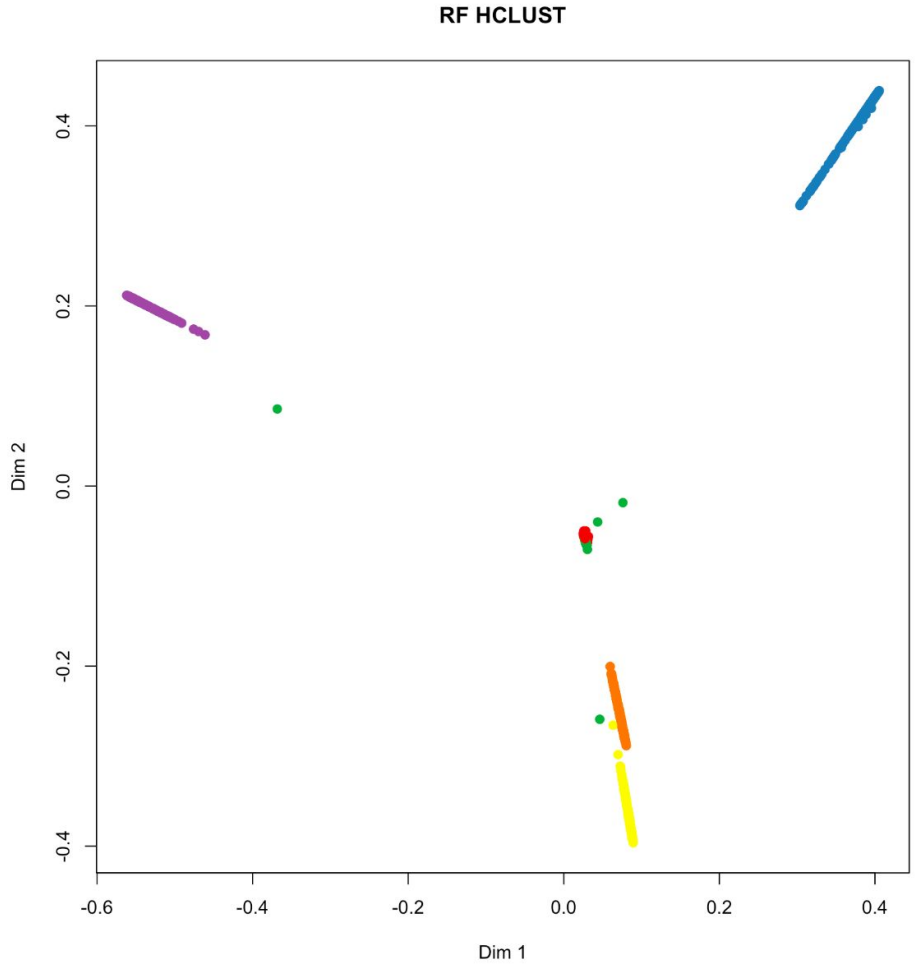

HCLUST
V2
V1
0.4
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
-0.2
-0.1
0.0
0.1
0.2

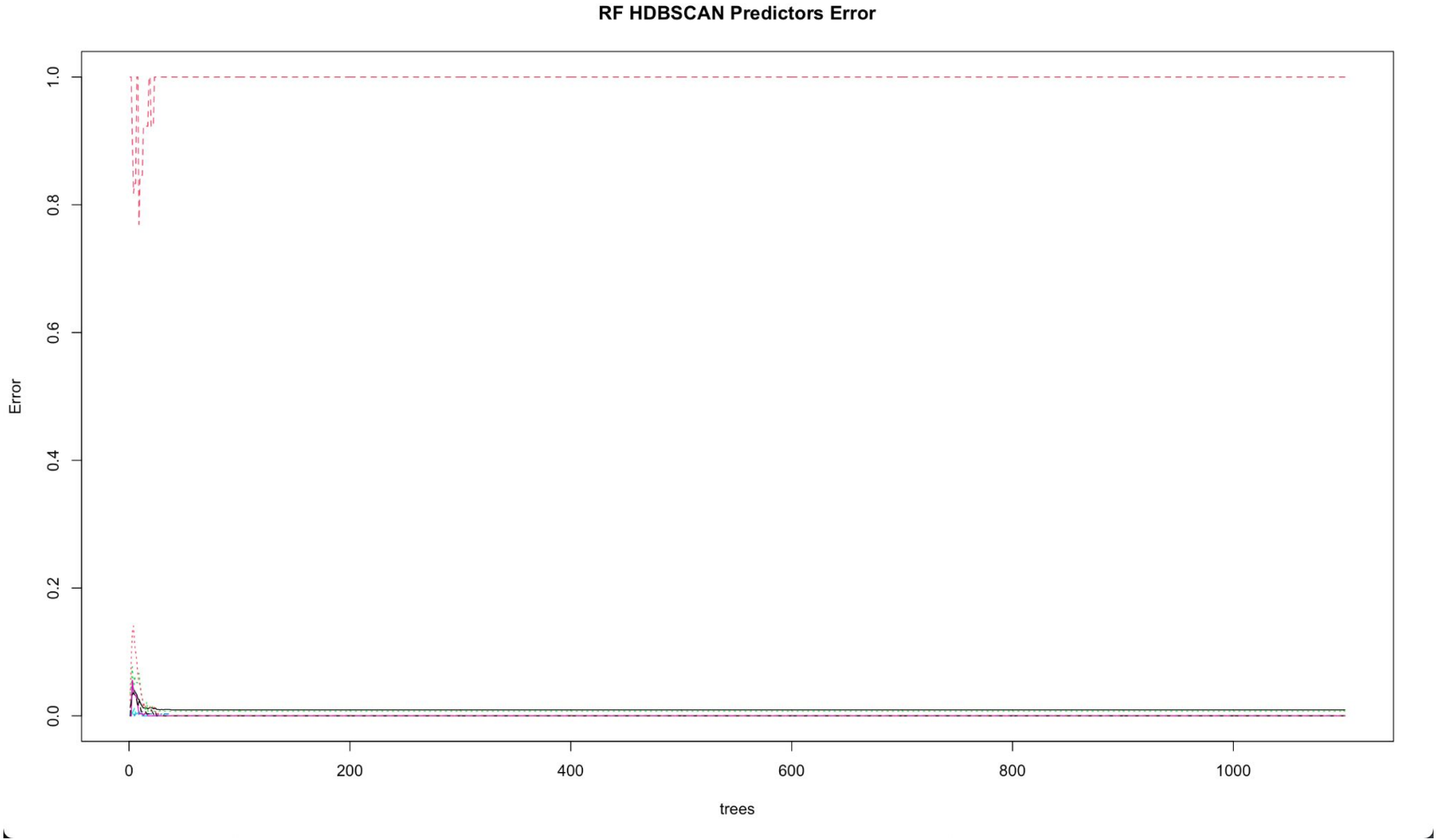

HDBSCAN
V2
V1
0.2
0.1
0.0
-0.1
-0.2
-0.2
-0.1
0.0
0.1
0.2

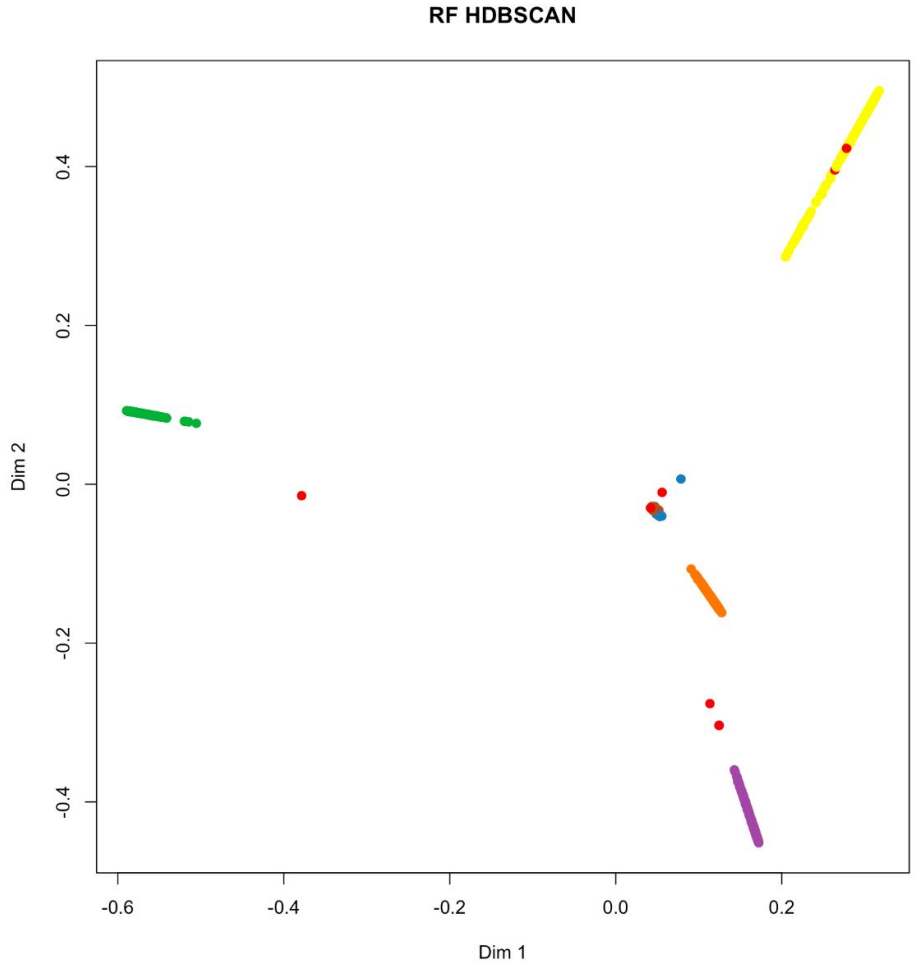

RF PAM Predictors Error
Error
0.00
0.02
0.04
0.06
0.08
0.10
0
200
400
600
800
1000
trees


RF PAM
Dim 2
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.4
-0.2
0.0
0.2
0.4
Dim 1


RF HCLUST Predictors Error
Error
0.00
0.02
0.04
0.06
0.08
0
200
400
600
800
1000
trees


RF HCLUST
Dim 2
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.4
-0.2
0.0
0.2
0.4
Dim 1


RF HDBSCAN Predictors Error
Error
0.0
0.2
0.4
0.6
0.8
1.0
0
200
400
600
800
1000
trees


RF HDBSCAN
Dim 2
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.4
-0.2
0.0
0.2
Dim 1

```
> RF.EARN[["importance"]]
                                          %IncMSE IncNodePurity
GPA                                     4880.2691     141039235
Number_Of_Professional_Connections     584015.2996    2057388875
Major                                 8809639.8968   42349063726
Graduation_Year                       -204.6793       81038460
Number_Of_Credits                     -2150.7916      52026449
Number_Of_Parking_Tickets               218.2527      29991946
> 
```

So, the importance of the variables in predicting the earnings are in this order based on IncNodePurity: Major => Connections => GPA => Graduation year => Credits => Parking Tickets

	ERROR	PAM	HCLUST	HDBSCAN
1	OOB	0.2	0.2	0.9333333
2	TEST	0.2	0.2	0.6000000

PAM and HCLUST had the lowest test set error rates of 0.2%. We will examine the medoids of PAM next.

	GPA	Number_Of_Professional_Connections	Earnings	Graduation_Year	Number_Of_Credits	Number_Of_Parking_Tickets	Major	PAM	HCLUST	HDBSCAN
1	2.65	10	10265.00	1992	122	1	Buisness	1	1	6
2	2.60	10	10254.77	1987	122	1	Humanities	2	2	5
3	2.61	11	5135.74	1986	121	1	Other	3	3	1
4	2.57	14	13258.34	1989	122	1	Vocational	4	4	2
5	2.52	7	9748.50	1987	122	1	STEM	5	5	4
6	2.37	7	11761.69	1990	121	1	Professional	6	6	3

These are the medoids of PAM and their respective cluster assignments

```
> RF.EARN

Call:
 randomForest(x = Data.Original[, .SD, .SDcols = !c("Earnings")],      y = Data.Original[, Earnings],
 ntree = 2777, mtry = 4, importance = TRUE,      proximity = TRUE, keep.forest = TRUE)
               Type of random forest: regression
                     Number of trees: 2777
No. of variables tried at each split: 4

          Mean of squared residuals: 16869.37
                    % Var explained: 99.62
```

MSE = 16869.37
% of the Variance explained = 99.62%