# FINELY TUNED PRESENTS: THE HIGGS PROJECT

CHRISTIAN HOLMES
WILL BARTLETT

# MISSING DATA

- Described by PRI_jet_num:
  - Missing when PRI_jet_num is 0: PRI_jet_leading_pt, PRI_jet_leading_eta, PRI_jet_leading_phi.
  - Missing when PRI_jet_num is 0 or 1: DER_deltaeta_jet_jet, DER_mass_jet_jet, DER_prodeta_jet_jet, DER_lep_eta_centrality, PRI_jet_subleading_pt, PRI_jet_subleading_eta, PRI_jet_subleading_phi.
- Randomly missing:
  - DER_mass_MMC (solved this by random imputation)

# WWCD (WHAT WOULD CHRIS DO?)

- 70.8% of rows have missing data!

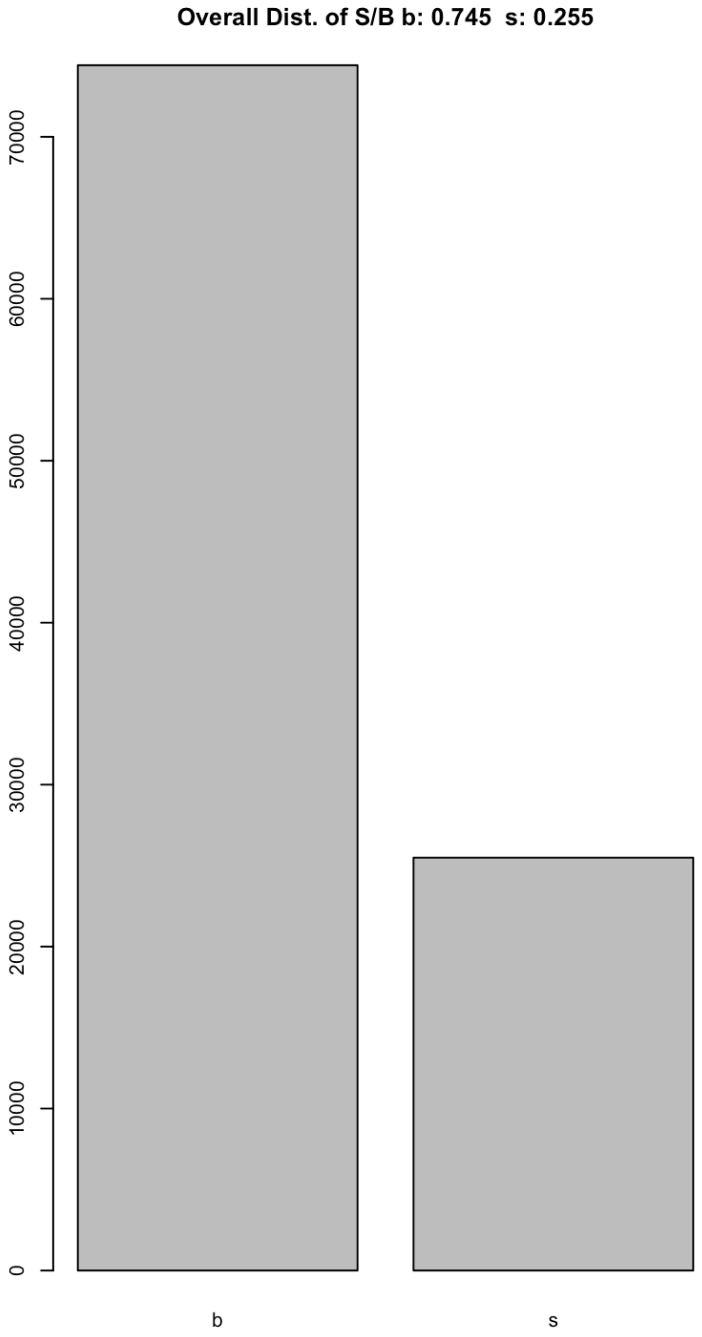
| | PRI_jet_num | sum(is.na(PRI_jet_leading_pt)) |
|---|---|---|
| | <int> | <int> |
| 1 | 0 | 99913 |
| 2 | 1 | 0 |
| 3 | 2 | 0 |
| 4 | 3 | 0 |

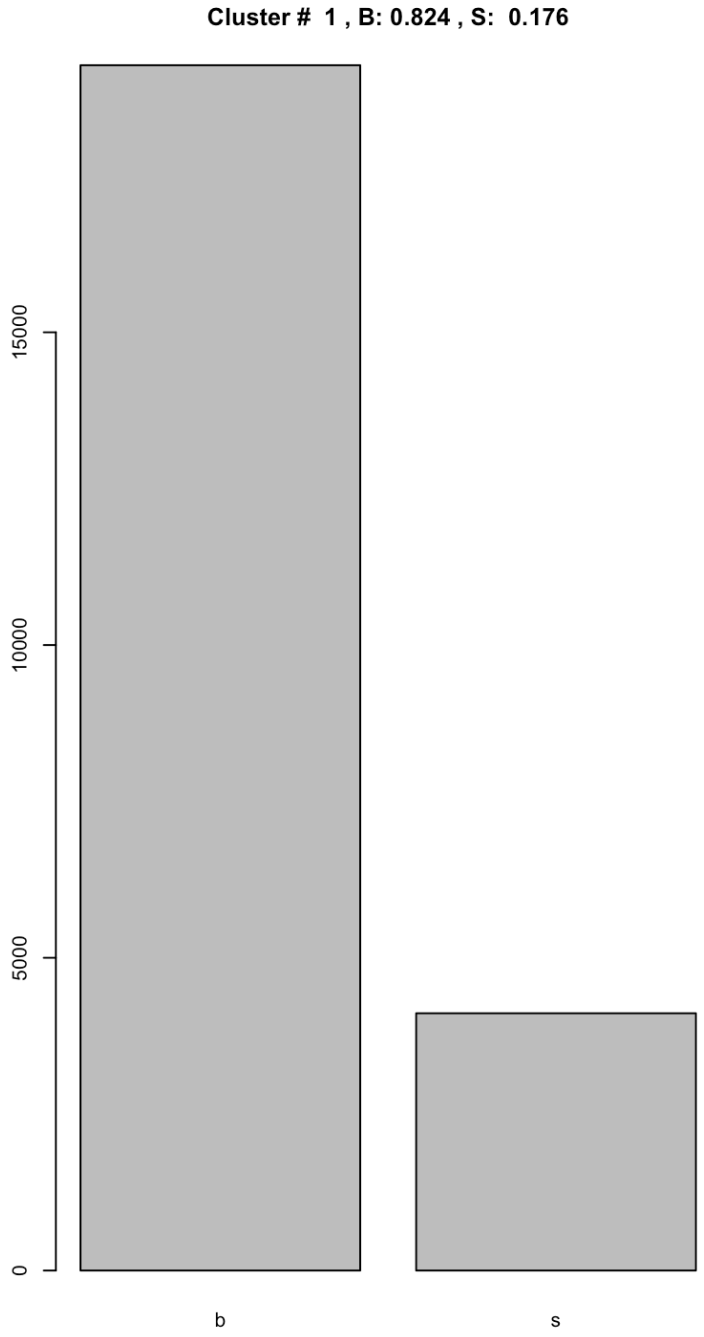
| | PRI_jet_num | sum(is.na(DER_deltaeta_jet_jet)) |
|---|---|---|
| | <int> | <int> |
| 1 | 0 | 99913 |
| 2 | 1 | 77544 |
| 3 | 2 | 0 |
| 4 | 3 | 0 |

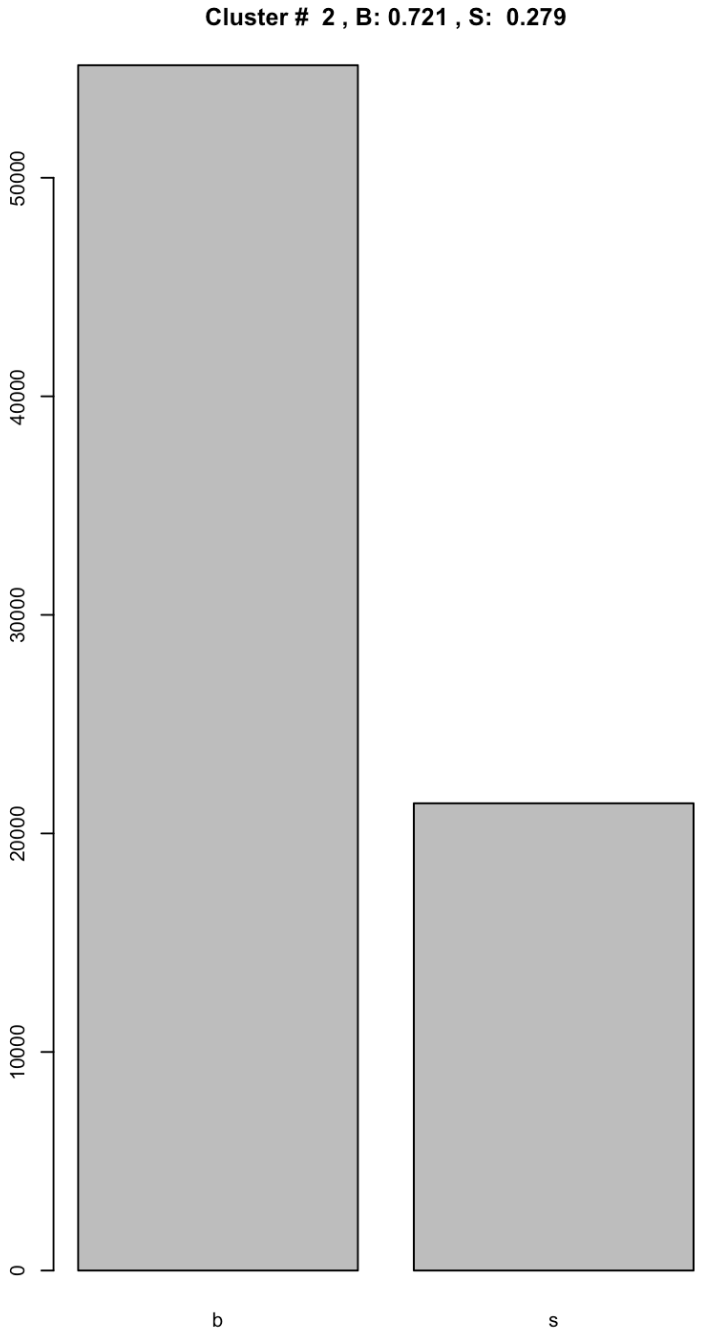
# EXPLORATORY ANALYSIS

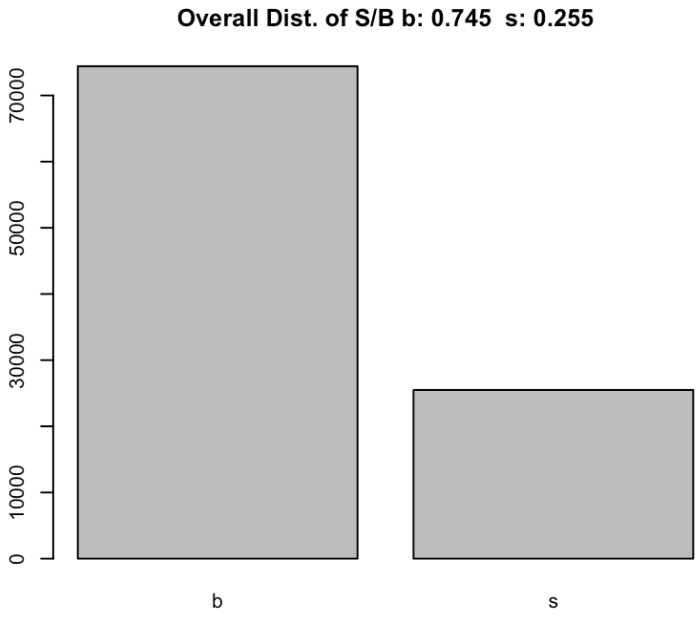
DISTRIBUTION OF S/B IN K MEANS CLUSTERS

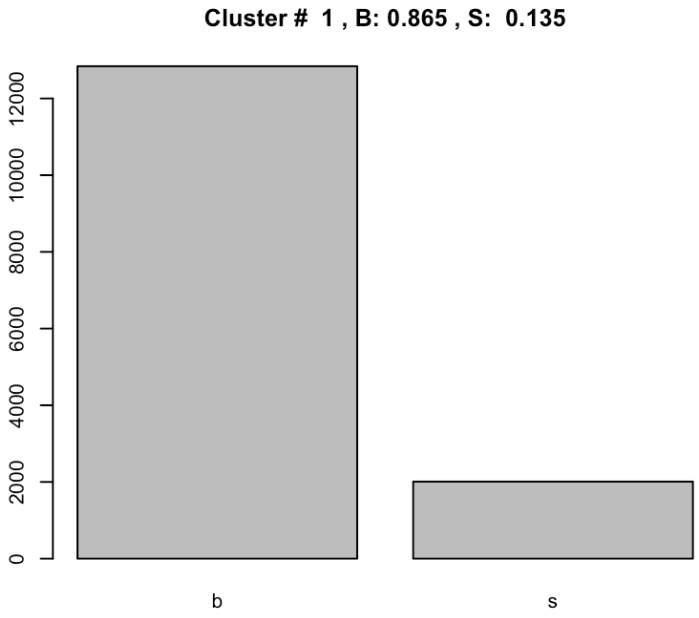
PRI_JET_NUM==0

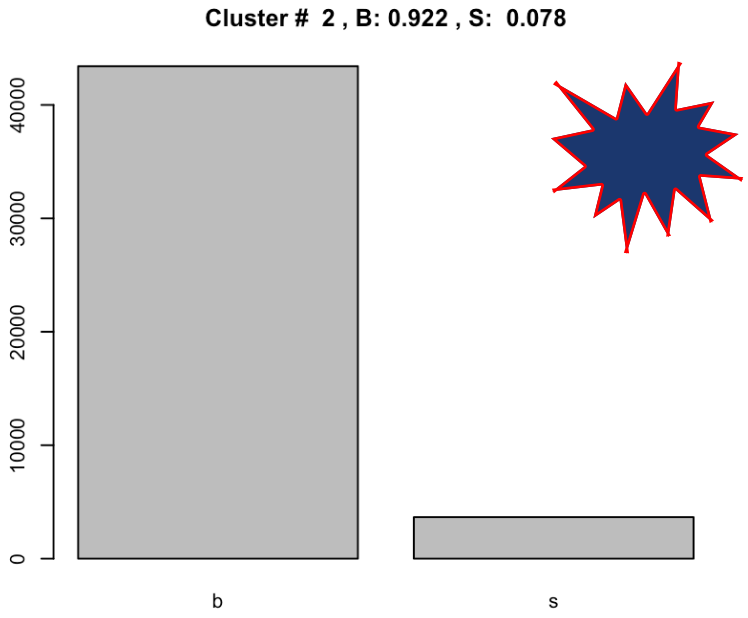

Overall Dist. of S/B b: 0.745 s: 0.255
0
10000
20000
30000
40000
50000
60000
70000
b
s

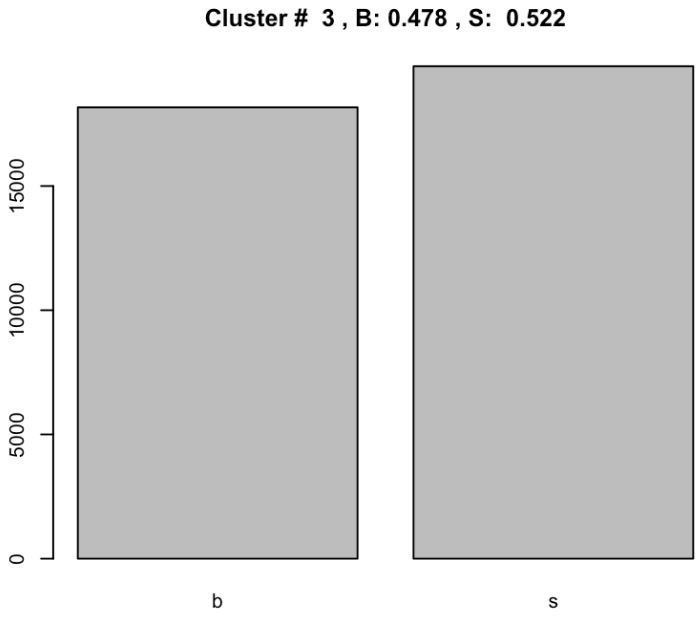

Cluster # 1 , B: 0.824 , S: 0.176
0
5000
10000
15000
b
s

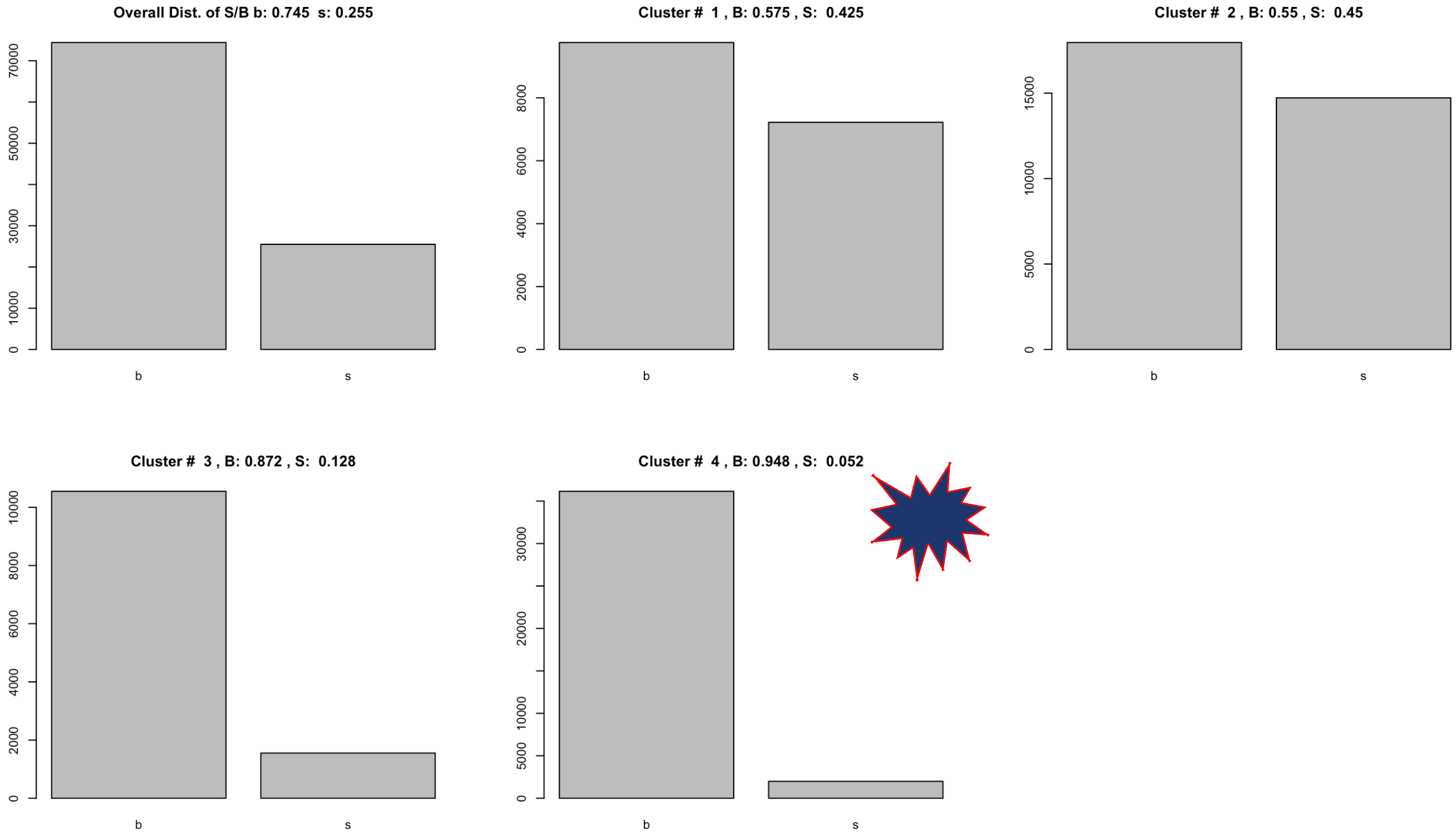

Cluster # 2 , B: 0.721 , S: 0.279
0
10000
20000
30000
40000
50000
b
s

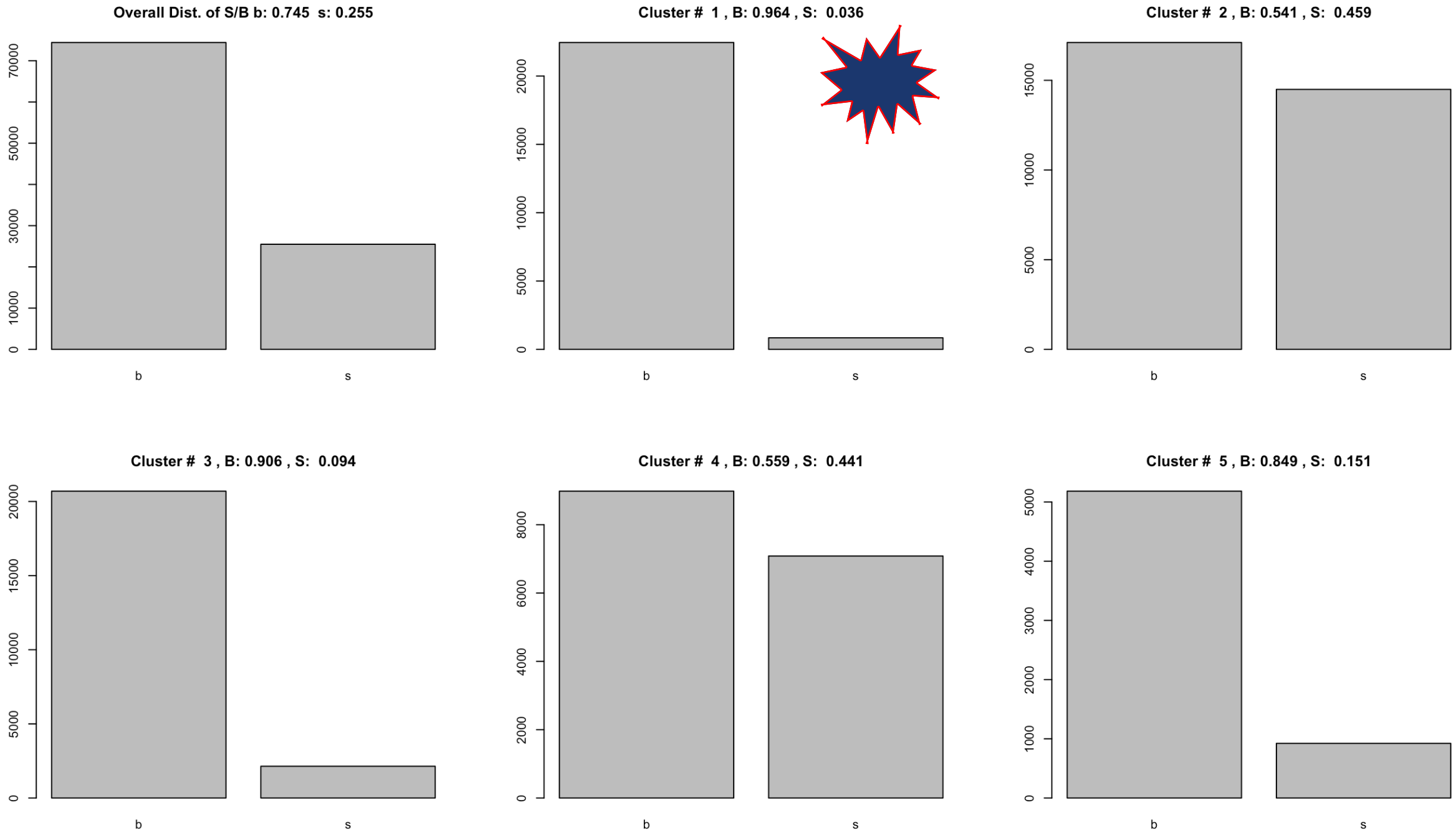

Overall Dist. of S/B b: 0.745 s: 0.255
0
10000
30000
50000
70000
b
s

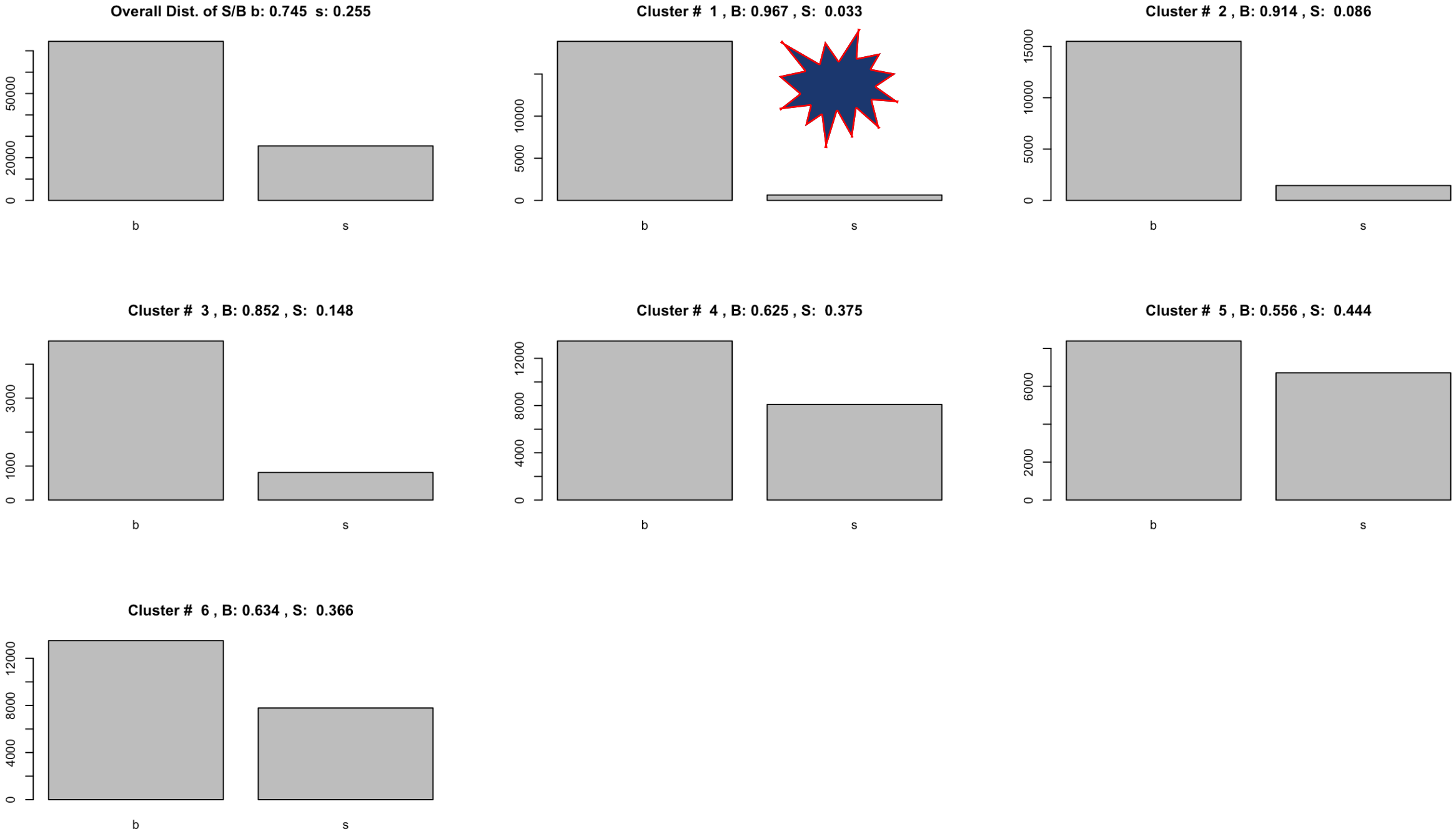

Cluster # 1 , B: 0.865 , S: 0.135
0
2000
4000
6000
8000
10000
12000
b
s

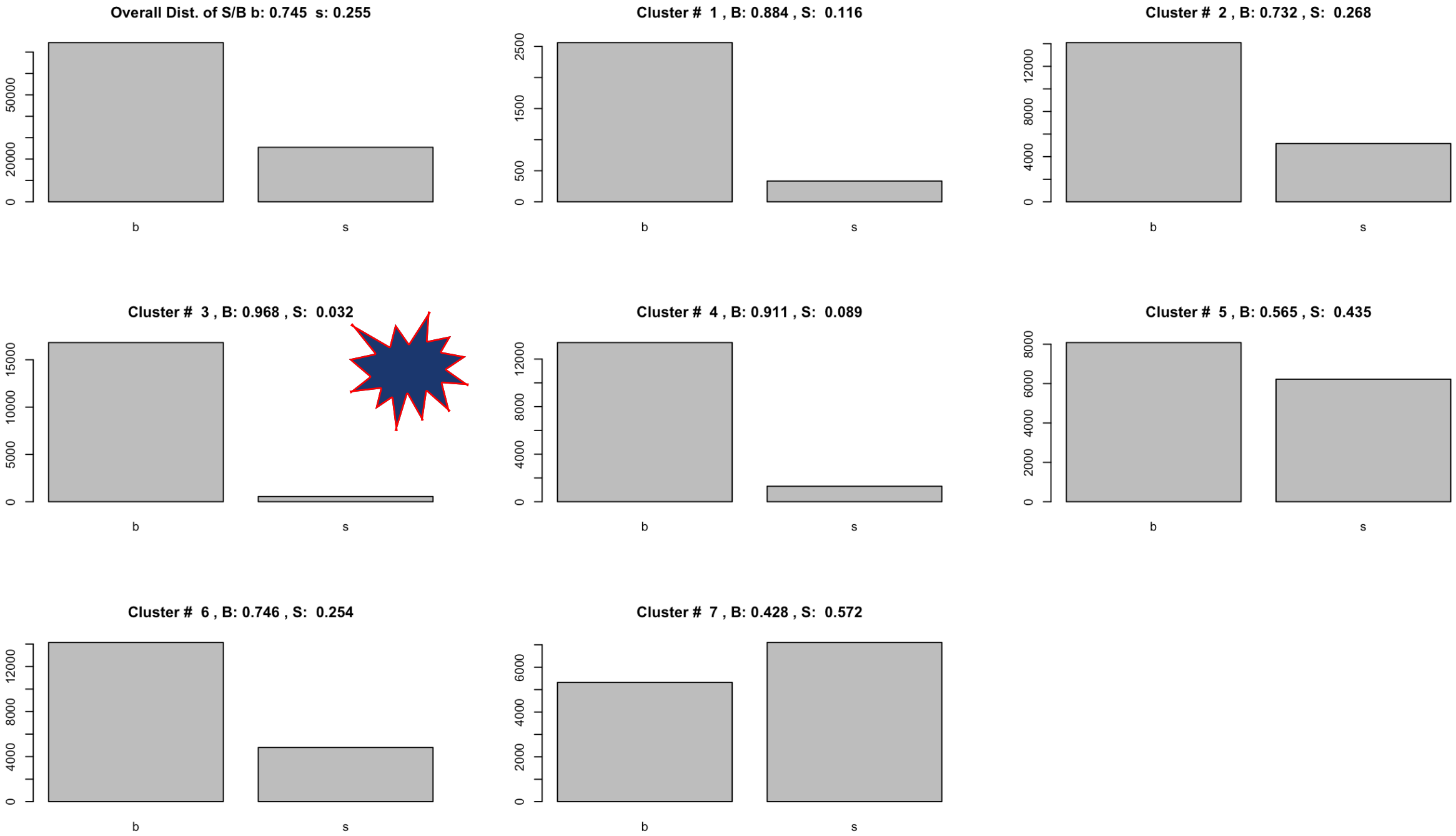

Cluster # 2 , B: 0.922 , S: 0.078
0
10000
20000
30000
40000
b
s

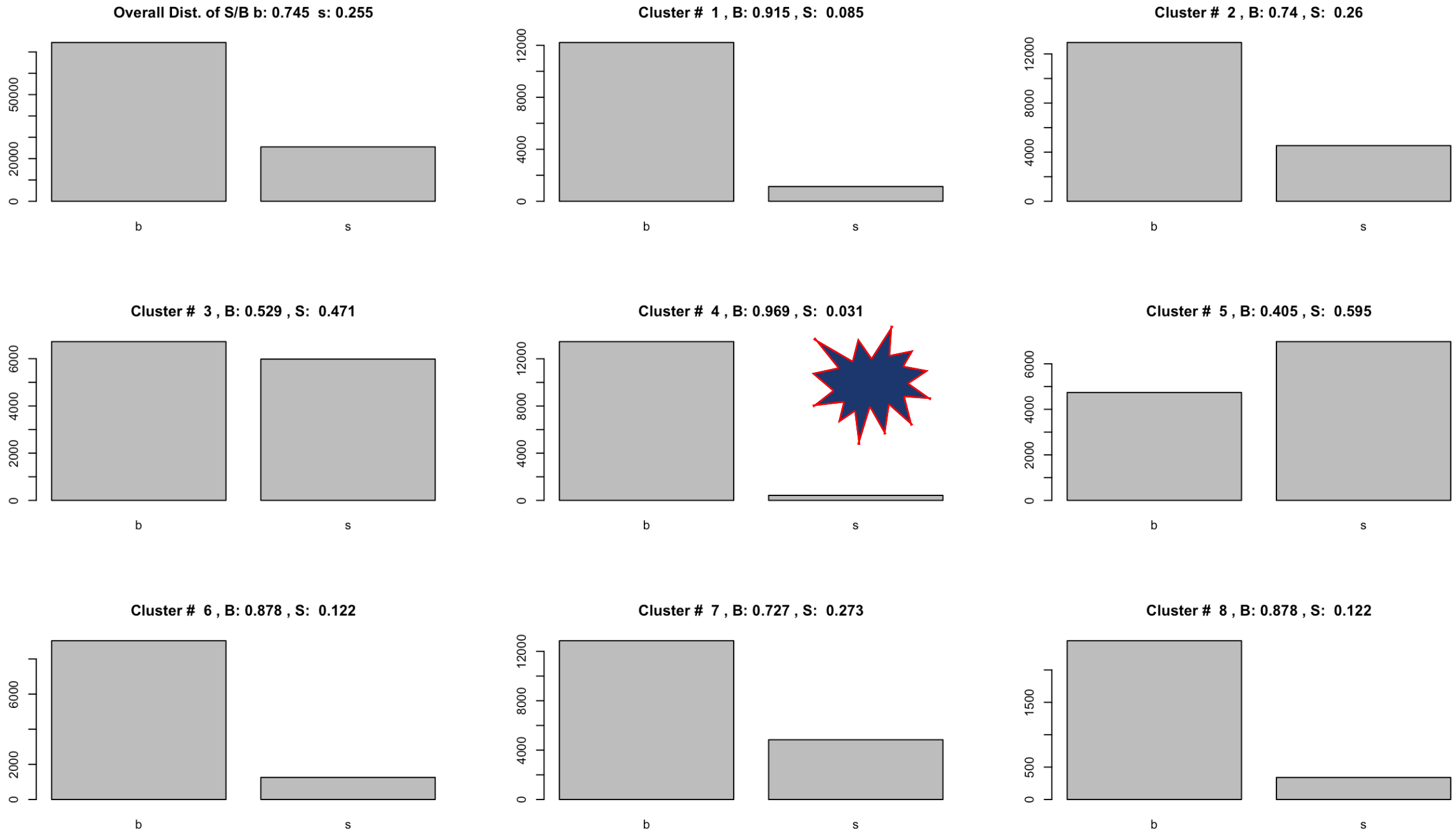

Cluster # 3 , B: 0.478 , S: 0.522
0
5000
10000
15000
b
s

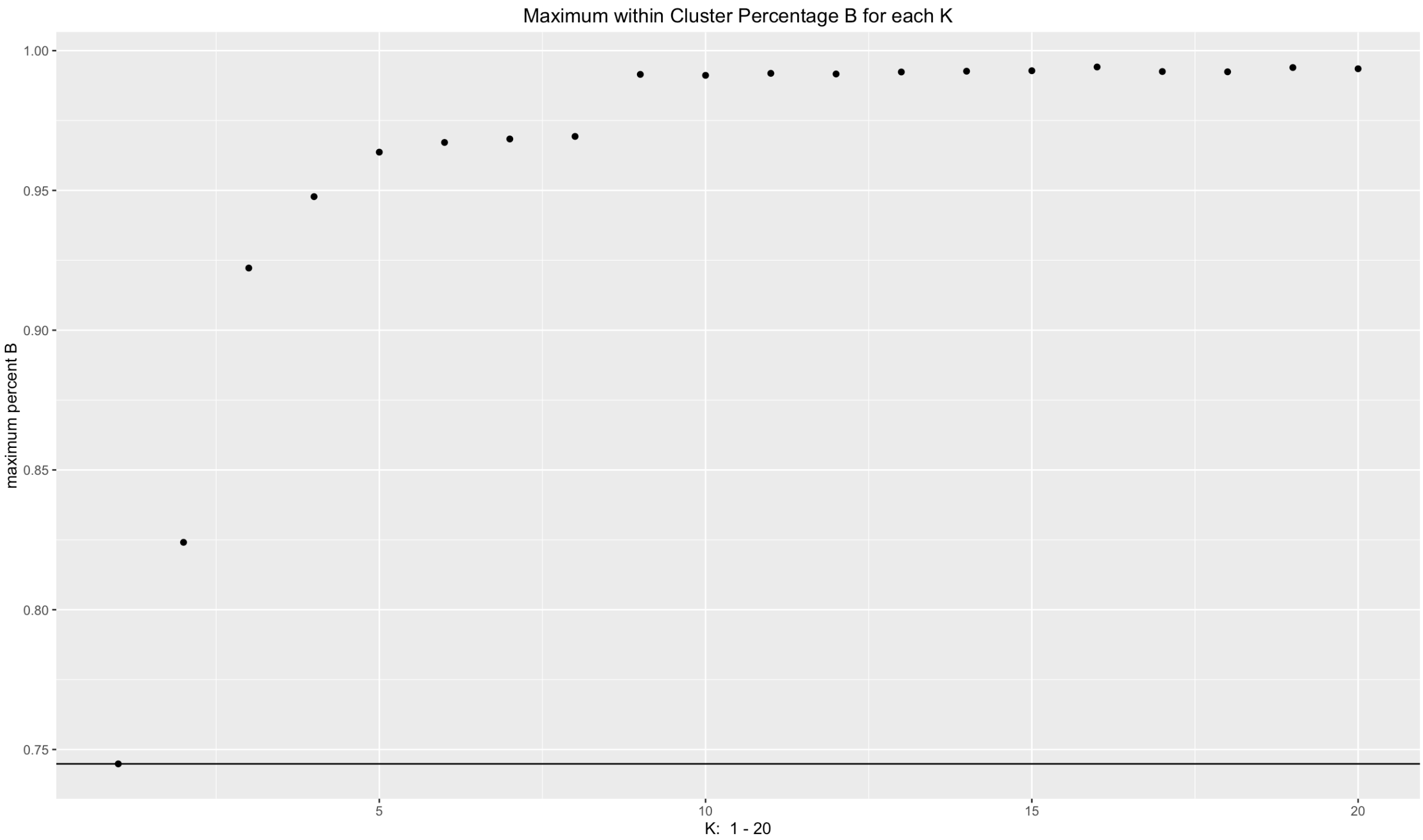

Overall Dist. of S/B b: 0.745 s: 0.255
0
10000
30000
50000
70000
b
s
Cluster # 1 , B: 0.575 , S: 0.425
0
2000
4000
6000
8000
b
s
Cluster # 2 , B: 0.55 , S: 0.45
0
5000
10000
15000
b
s
Cluster # 3 , B: 0.872 , S: 0.128
0
2000
4000
6000
8000
10000
b
s
Cluster # 4 , B: 0.948 , S: 0.052
0
5000
10000
20000
30000
b
s

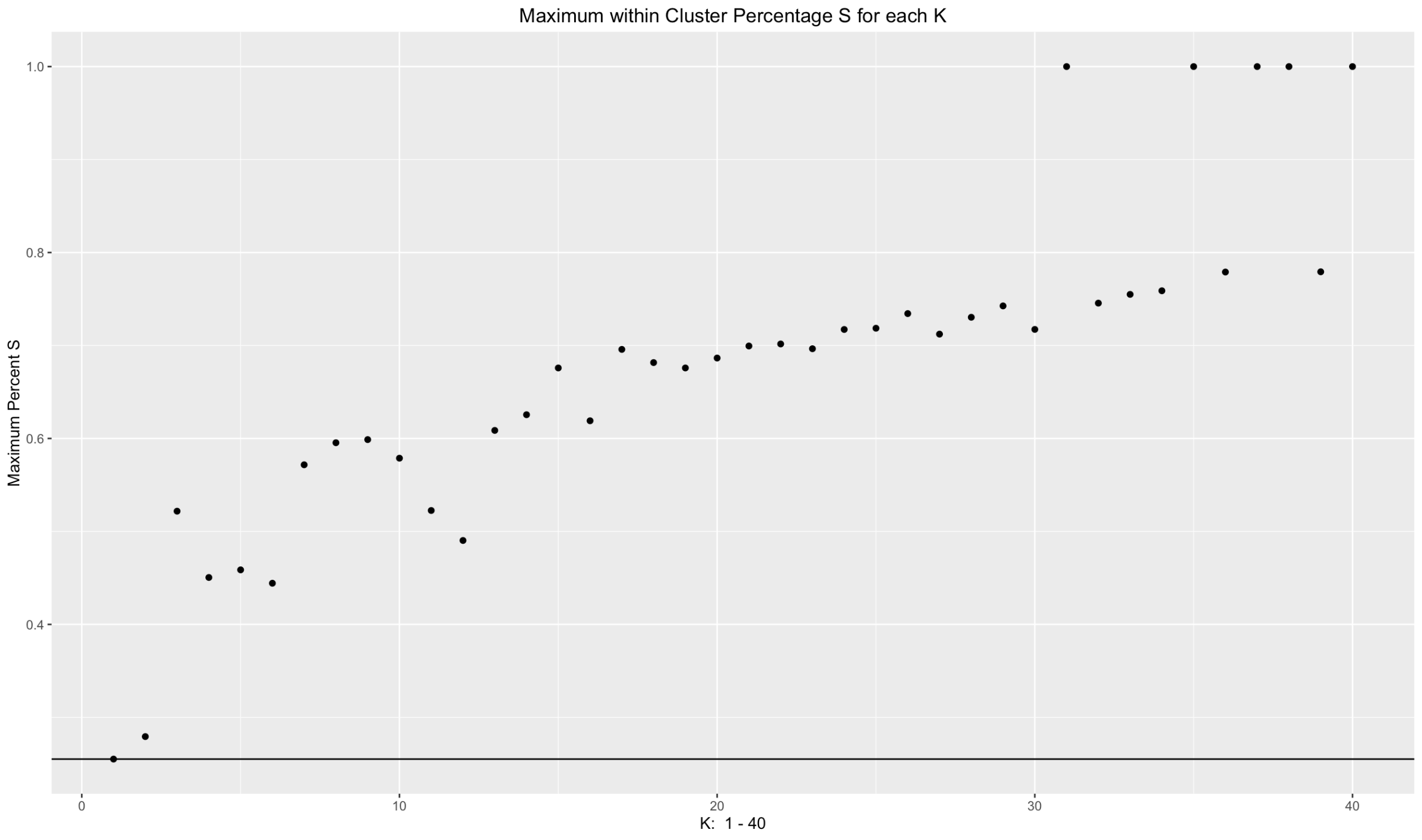

Overall Dist. of S/B b: 0.745 s: 0.255
Cluster # 1 , B: 0.964 , S: 0.036
Cluster # 2 , B: 0.541 , S: 0.459
Cluster # 3 , B: 0.906 , S: 0.094
Cluster # 4 , B: 0.559 , S: 0.441
Cluster # 5 , B: 0.849 , S: 0.151
b
s

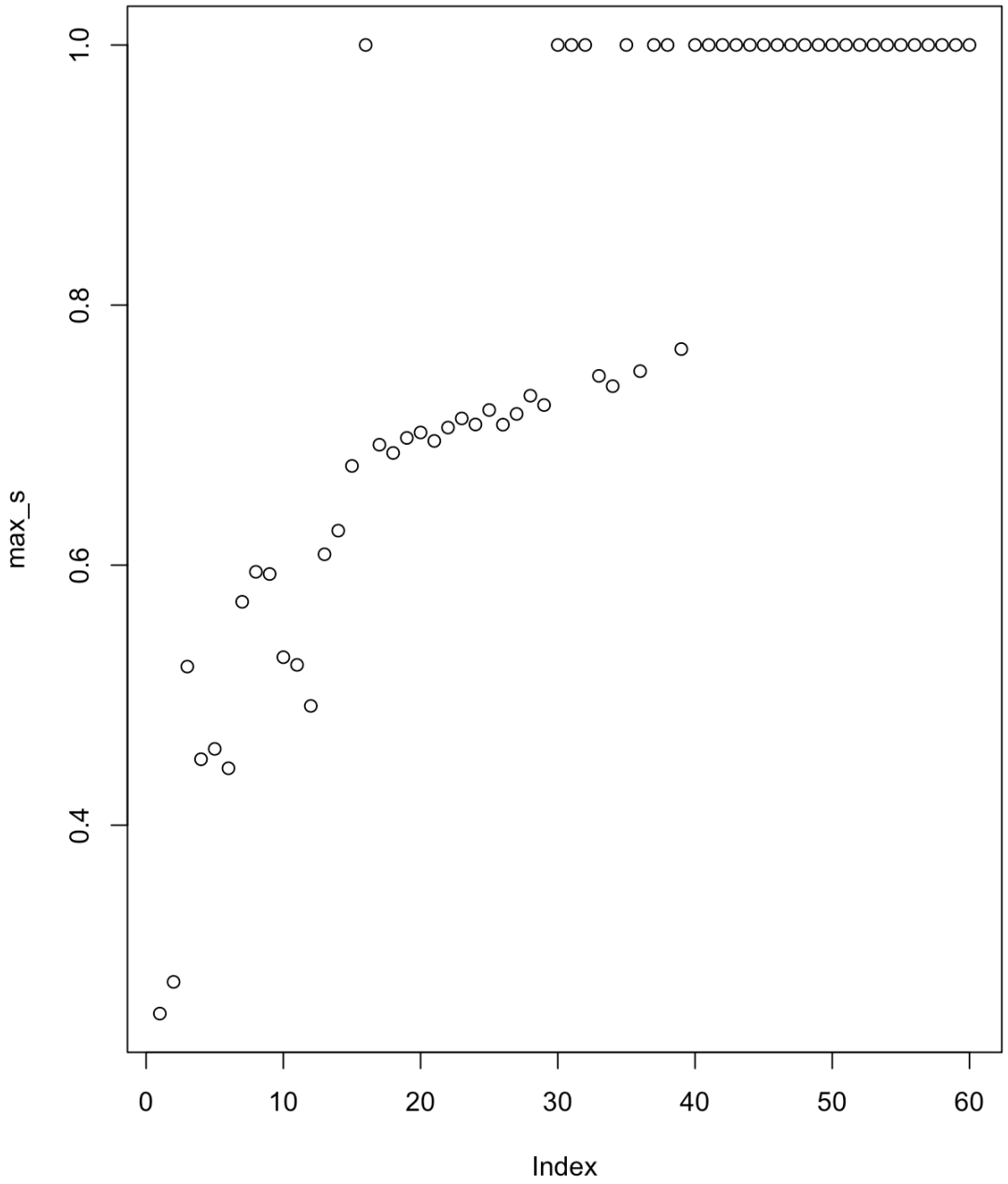

Overall Dist. of S/B b: 0.745 s: 0.255
0
20000
50000
b
s
Cluster # 1 , B: 0.967 , S: 0.033
0
5000
10000
b
s
Cluster # 2 , B: 0.914 , S: 0.086
0
5000
10000
15000
b
s
Cluster # 3 , B: 0.852 , S: 0.148
0
1000
3000
b
s
Cluster # 4 , B: 0.625 , S: 0.375
0
4000
8000
12000
b
s
Cluster # 5 , B: 0.556 , S: 0.444
0
2000
6000
b
s
Cluster # 6 , B: 0.634 , S: 0.366
0
4000
8000
12000
b
s

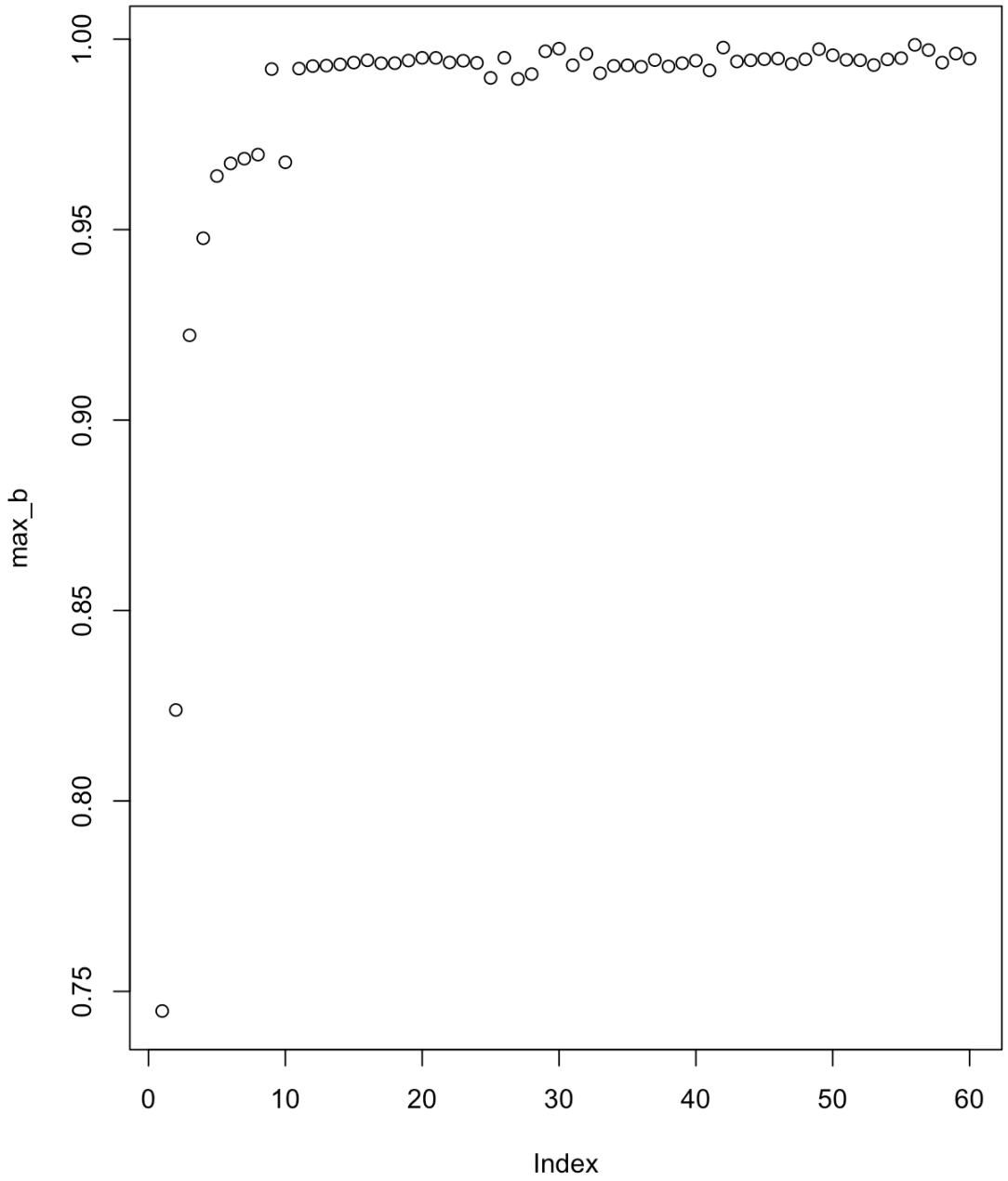

Overall Dist. of S/B b: 0.745 s: 0.255
Cluster # 1 , B: 0.884 , S: 0.116
Cluster # 2 , B: 0.732 , S: 0.268
Cluster # 3 , B: 0.968 , S: 0.032
Cluster # 4 , B: 0.911 , S: 0.089
Cluster # 5 , B: 0.565 , S: 0.435
Cluster # 6 , B: 0.746 , S: 0.254
Cluster # 7 , B: 0.428 , S: 0.572
b
s

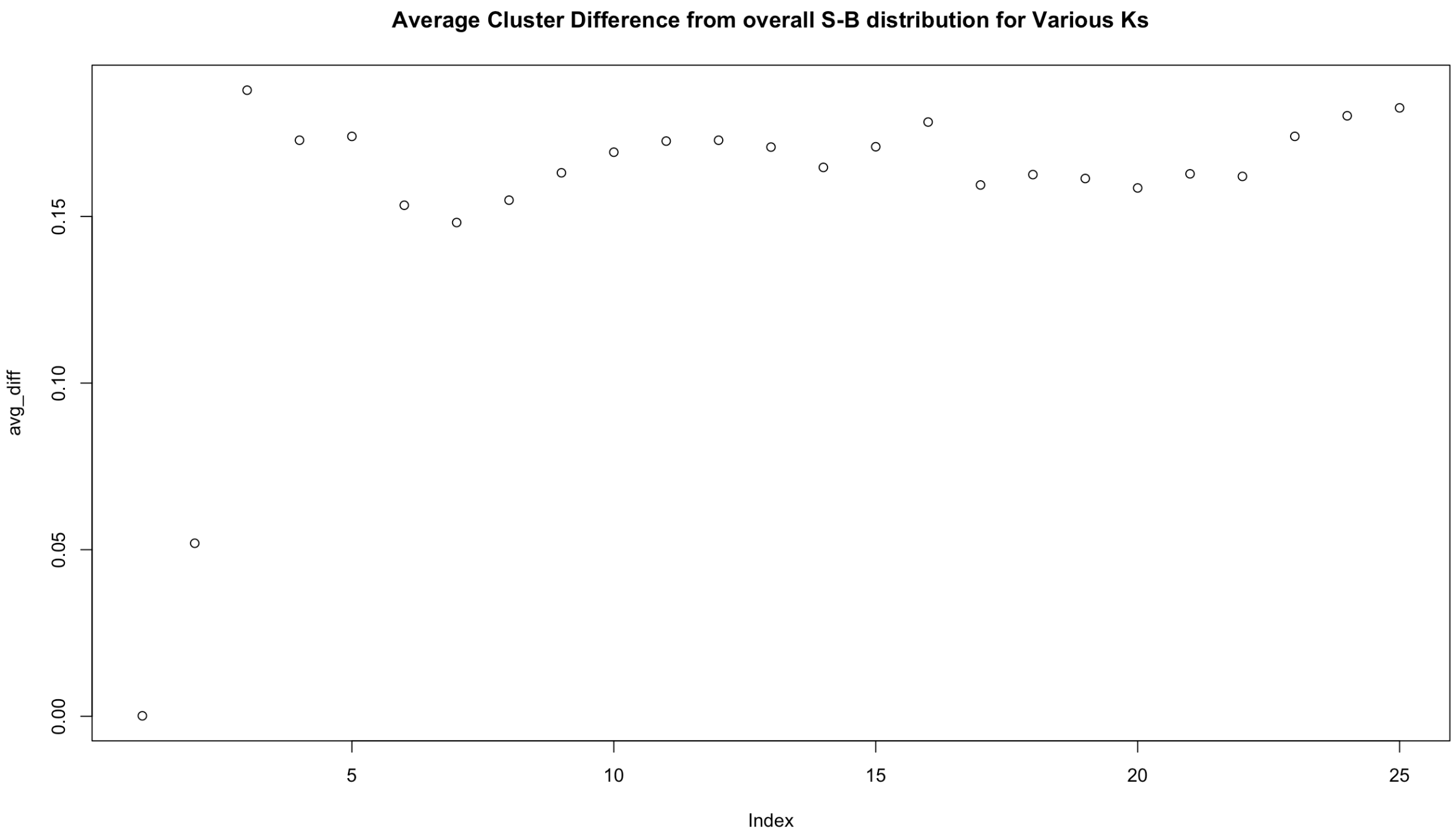

Overall Dist. of S/B b: 0.745 s: 0.255
Cluster # 1 , B: 0.915 , S: 0.085
Cluster # 2 , B: 0.74 , S: 0.26
Cluster # 3 , B: 0.529 , S: 0.471
Cluster # 4 , B: 0.969 , S: 0.031
Cluster # 5 , B: 0.405 , S: 0.595
Cluster # 6 , B: 0.878 , S: 0.122
Cluster # 7 , B: 0.727 , S: 0.273
Cluster # 8 , B: 0.878 , S: 0.122
b
s


Maximum within Cluster Percentage B for each K
maximum percent B
1.00
0.95
0.90
0.85
0.80
0.75
5
10
15
20
K: 1 - 20


Maximum within Cluster Percentage S for each K
Maximum Percent S
1.0
0.8
0.6
0.4
0
10
20
30
40
K: 1 - 40


max_s
1.0
0.8
0.6
0.4
0
10
20
30
40
50
60
Index


max_b
1.00
0.95
0.90
0.85
0.80
0.75
0
10
20
30
40
50
60
Index


Average Cluster Difference from overall S-B distribution for Various Ks
avg_diff
0.00
0.05
0.10
0.15
5
10
15
20
25
Index

# BUT, HOW POWERFUL ARE THESE CLUSTER FACTOR VARIABLES?

## LOGISTIC REGRESSION

Logistic Regression Trials with Added Cluster Factor Variables for JET_PRI_NUM = 0

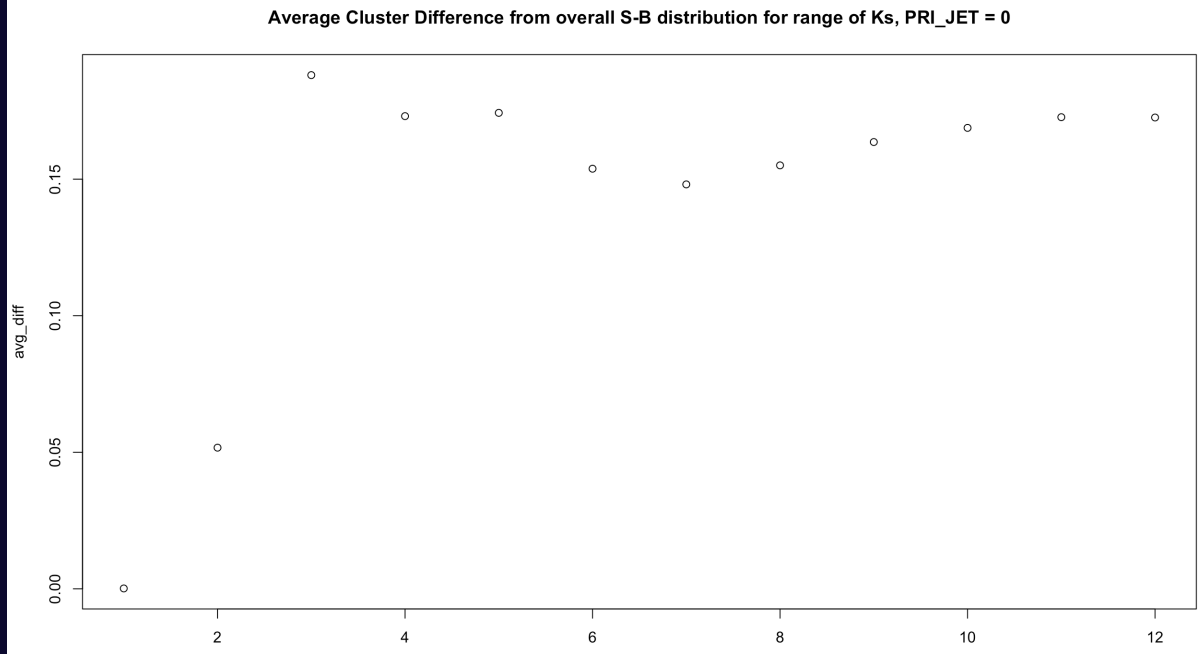
| Factor Variables for K = | BIC | AIC | Performance Over Majority Guess (74.49%) |
|---|---|---|---|
| No Cluster Factor Variables | 78422.64 | 78241.91 | 8.06802 |
| 3 | 76165.01 | 75965.26 | 8.59047 |
| 2:8 | 75373.72 | 75040.80 | 8.74461 |
| 2:15 | 75261.75 | 74833.71 | 8.865713 |

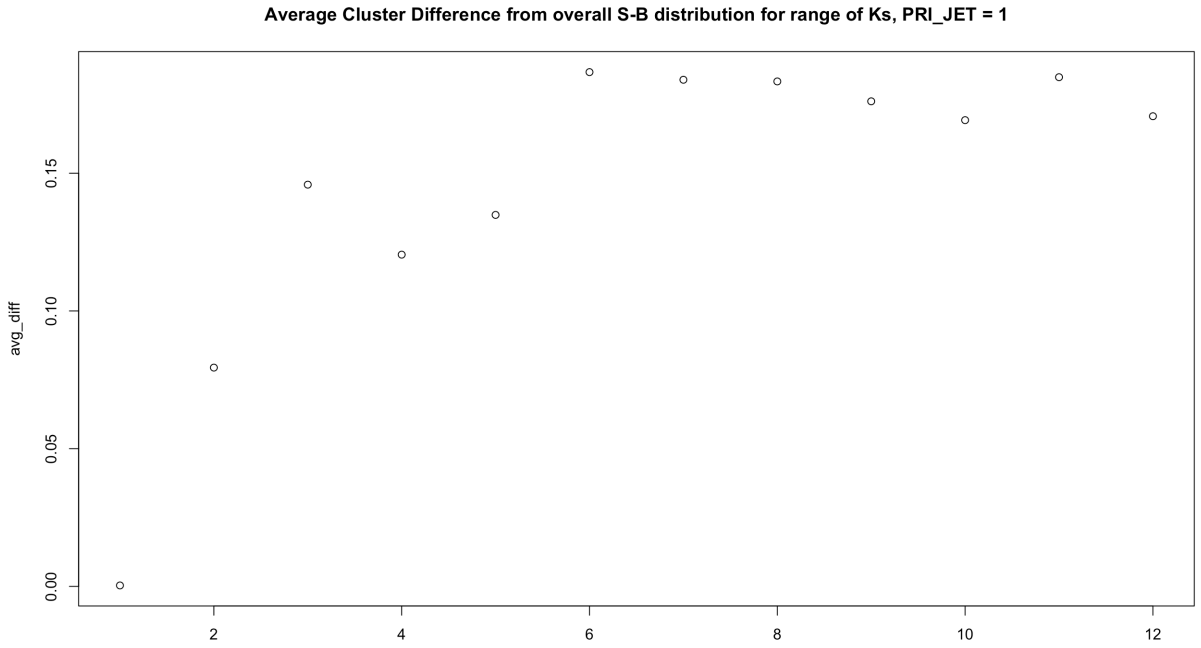
Logistic Regression Trials with Added Cluster Factor Variables
for JET_PRI_NUM = 0

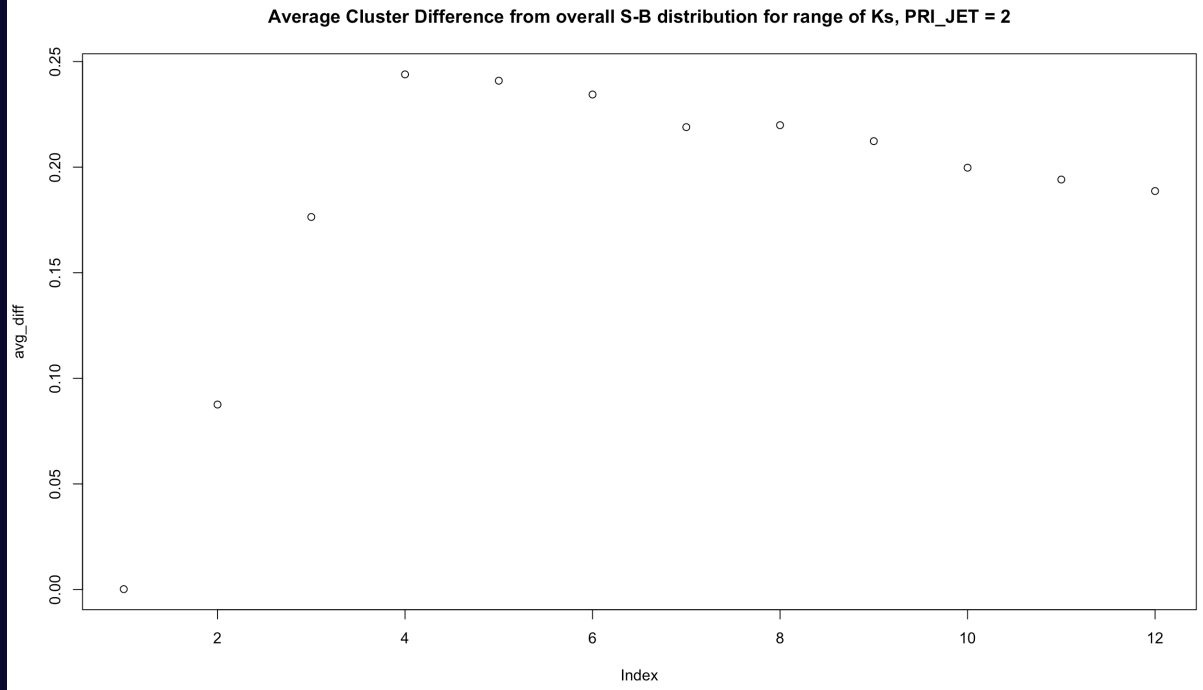
| Factor Variables for K = | BIC | AIC | Performance Over Majority Guess (74.49%) |
|---|---|---|---|
| No Cluster Factor Variables | 78422.64 | 78241.91 | 8.06802 |
| 3 | 76165.01 | 75965.26 | 8.59047 |
| 2:8 | 75373.72 | 75040.80 | 8.74461 |
| 2:8 & no PRI's**** | 75419.73 | 75172.42 | 8.69957 |
| 2:15 | 75261.75 | 74833.71 | 8.865713 |

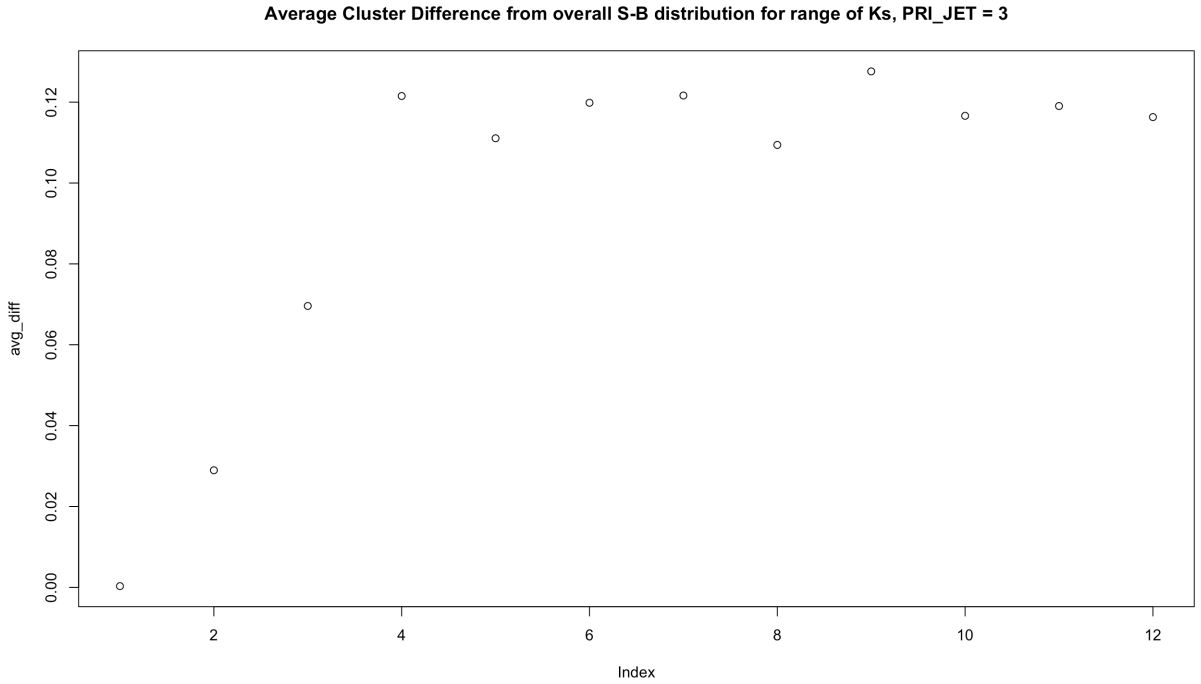
- Getting rid of ALL PRIs hurt the model a little bit, but not even close to as much as not having any factor variables.

# WHAT ABOUT THE OTHER SPLITS?


Average Cluster Difference from overall S-B distribution for range of Ks, PRI_JET = 0
avg_diff
0.00
0.05
0.10
0.15
2
4
6
8
10
12


Average Cluster Difference from overall S-B distribution for range of Ks, PRI_JET = 1
avg_diff
0.00
0.05
0.10
0.15
2
4
6
8
10
12


Average Cluster Difference from overall S-B distribution for range of Ks, PRI_JET = 2
avg_diff
0.00
0.05
0.10
0.15
0.20
0.25
2
4
6
8
10
12
Index


Average Cluster Difference from overall S-B distribution for range of Ks, PRI_JET = 3
avg_diff
0.00
0.02
0.04
0.06
0.08
0.10
0.12
2
4
6
8
10
12
Index

Logistic Regression Trials with Added Cluster Factor Variables for JET_PRI_NUM = 2

| Factor Variables for K = | BIC | AIC | Performance Over Majority Guess (.511%) |
|---|---|---|---|
| No Cluster Factor Variables | 52258.43 | 52002.44 | 22.94408 |
| 2 | 52263.26 | 51998.44 | 22.95996 |
| 4 | 49410.05 | 49127.58 | 25.73096 |
| 2:4 | 49321.91 | 49012.96 | 25.83418 |
| 2:8 | 48579.17 | 48217.25 | 26.67381 |

# GBM NOT SPLIT VS SPLIT

| | Score | Position |
|---|---|---|
| Not Split | 1.16 | 1639 |
| Split | 2.02 | 1495 |

# XGBOOST NOT SPLIT VS SPLIT

| | Score | Position |
|---|---|---|
| Not Split | 2.36 | 1371 |
| Split | 2.83 | 1184 |

# XGBOOST WITH WILL'S CLUSTER FACTOR VARIABLES

- …didn't quite get there…

# FUTURE AREAS OF INTEREST

- Add many factor variables to each split rather than just one.
  - Logistic BIC, AIC, and model accuracy continued to improve with each addition of KMeans factor variables—unclear when these metrics would indicate too many variables.
- Figure out the best combinations of factor variables.