

Exercise 1

1. The first feature set is `c("length", "diameter", "height")`
I choose the $k=60$ to do the comparison

contingency tables:

	actual		
predicted	young	adult	old
young	150	37	12
adult	47	184	79
old	6	25	36

Accuracy of feature set 1: 0.6424

2. The second feature set is `c("whole_weight", "shucked_weight", "viscera_weight", "shell_weight")`

contingency tables:

	actual		
predicted	young	adult	old
young	155	36	8
adult	46	192	70
old	2	18	49

Accuracy of feature set 2: 0.6875

3. If we iterate with $k=1:100$ for training for feature set 1 and feature set 2. Then we get the following result:
Best accuracy of feature set 1: 0.6528 at $k=23$
Best accuracy of feature set 2: 0.6944 at $k=28$

Therefore we choose feature set 2 `c("whole_weight", "shucked_weight", "viscera_weight", "shell_weight")` as best performing feature subset for exercise 2

Exercise 2

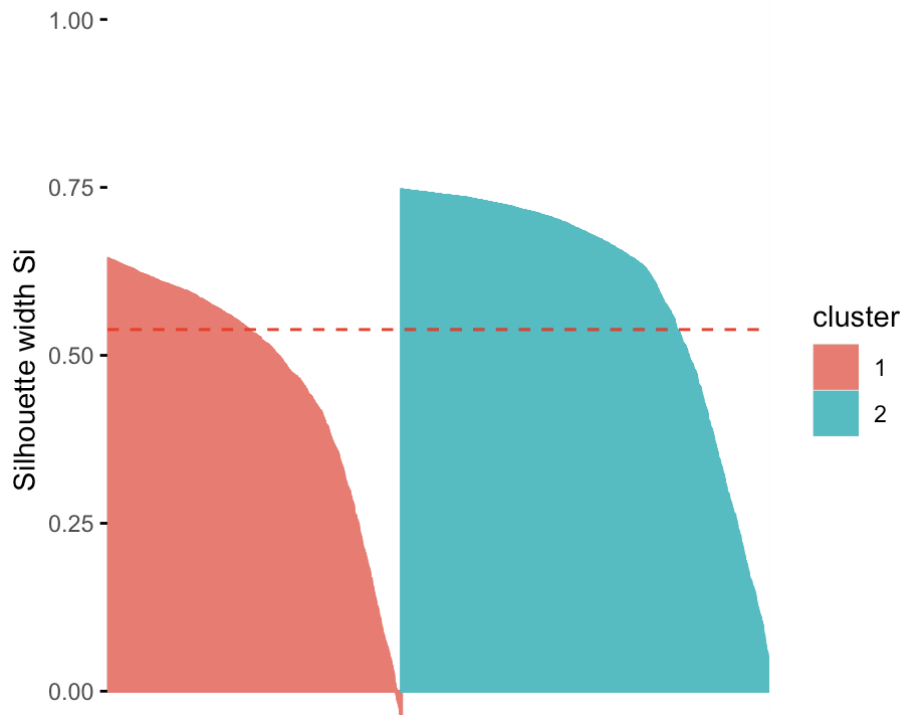
We look at k=2:10 for the dataset

Optimal number of clusters for K-Means with silhouette method is 2, so the rest is based on the result for k=2

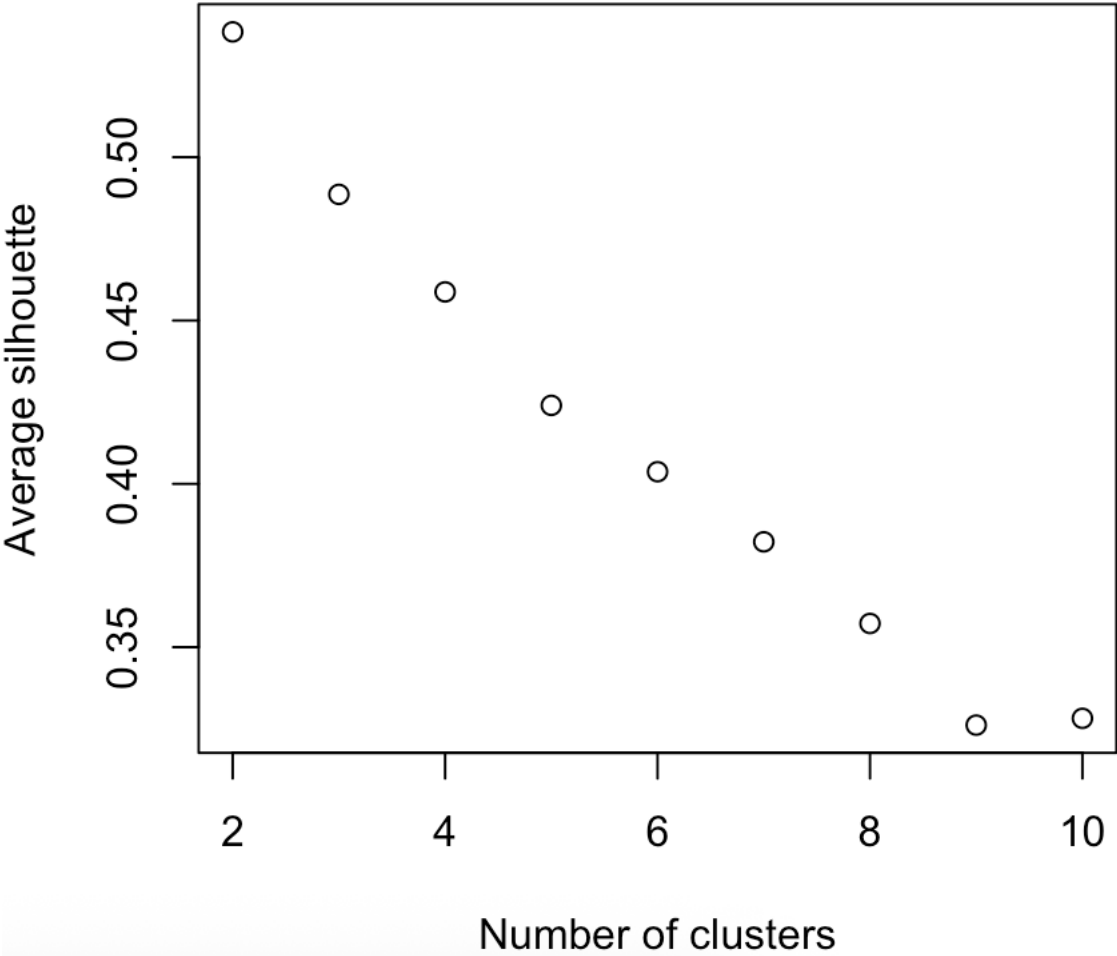
```
> fviz_silhouette(sil_km)
```

	cluster	size	ave.sil.width
1	1	1857	0.46
2	2	2319	0.60

Clusters silhouette plot
Average silhouette width: 0.54



Silhouette vs K for K-Means

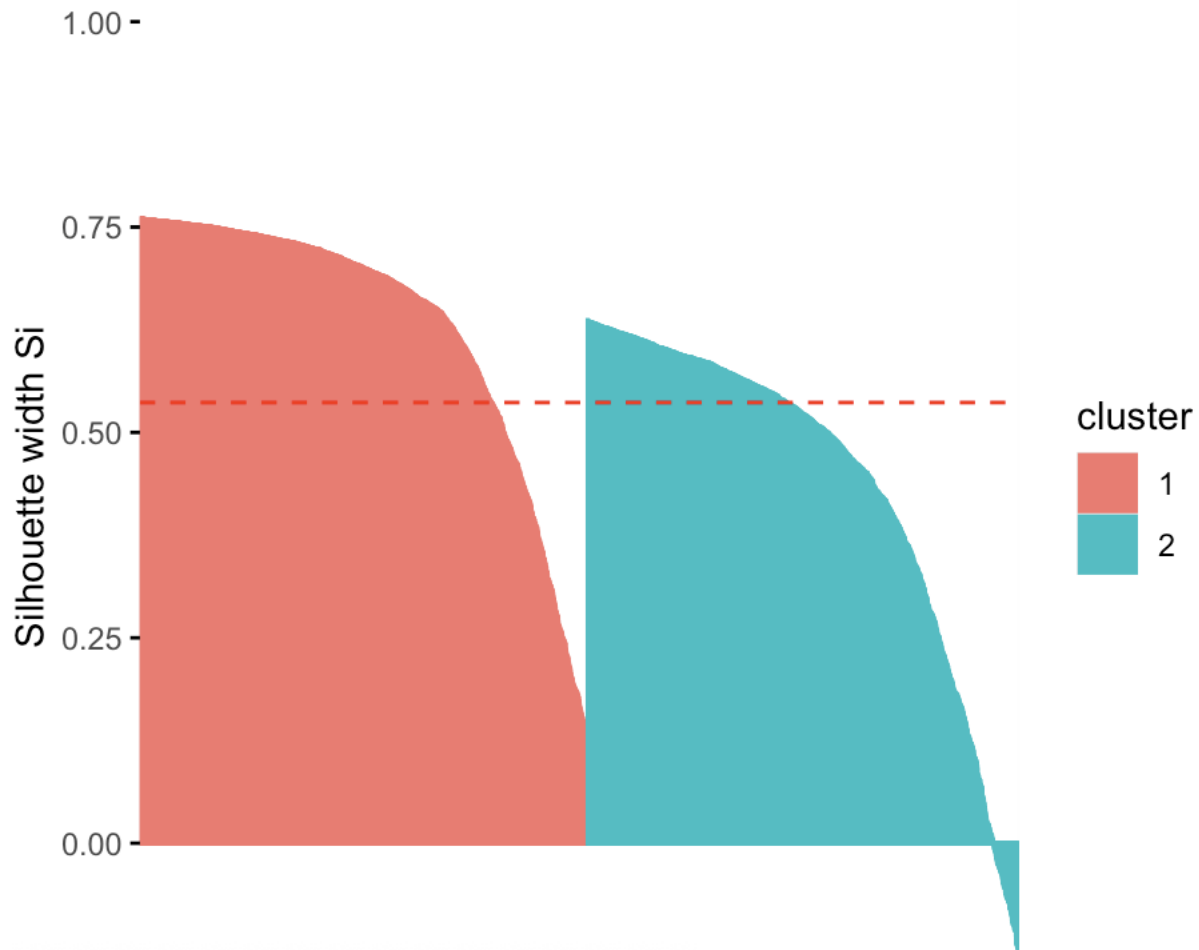


Optimal number of clusters for PAM with silhouette method is 2, and the rest is based on the result of k=2.

```
> fviz_silhouette(sil_pam)
```

	cluster	size	ave.sil.width
1	1	2127	0.63
2	2	2049	0.44

Clusters silhouette plot
Average silhouette width: 0.54



Silhouette vs K for PAM

