Question 4

Purpose

`clusterboot()` from package `fpc` is a useful function that resamples the original data at a given time using various methods such as bootstrap and subsetting. The Jaccard coefficient, the similarity measure between resampling clusters and clusters found by the clustering method, is calculated as the number of points in the intersection divided by the number of points in the union of these two sets. A Jaccard coefficient under 0.5 indicates the dissolution of a cluster when adding more points into the dataset. Then it evaluates the clusterwise stability of each cluster through the mean over these Jaccard similarities (Hennig, 2020).

Stability, an element of cluster validation, defines that "a meaningful valid cluster should not easily disappear if the data set is changed in a non-essential way" and is often a popular way to predict the true cluster number (Hennig, 2007). Moreover, a cluster's stability is heavily reliant on the homogeneity and separation of the clusters. When the stability value is < 0.6, the cluster is unstable, whereas to be stable, the value needs to be ≥ 0.75. Those clusters with values in between are considered to demonstrate a pattern, but there is uncertainty around which points should be grouped together (Hennig, 2020). `clusterboot()` also measures the clustering algorithm's stability as each algorithm can produce different stability values (Zumel & Mount, 2014).

General Steps of the Algorithm

1. Perform clustering method (e.g. K-means, hierarchical clustering) on the dataset.
2. Resample the data. The algorithm will perform clustering on the resampled datasets based on the clustering method chosen in step 1. The following are different resampling methods:
    2.1. Bootstrap: a number of observations from the original dataset are selected randomly to create a bootstrap dataset. Along with sampling, replacement is also performed to keep the sample size equal to the original dataset size.
    2.2. Subsetting: data is resampled by choosing a subset of observations from original data. The number of observations in each subset should be carefully considered to generate enough variation and avoid poor clustering results.
    2.3. Noise: a number of *m* points from original data are replaced by points drawn from a noise distribution. The noise distribution and *m* should be considered to create noise points to weaken clusters.

2.4. Jittering: noise is added to all points in the dataset. Jittering and bootstrap can be combined to avoid multiple points issues by adding noise in the bootstrap dataset.

3. For every single cluster in the original clustering, find the most similar cluster and record the Jaccard coefficient.
4. Repeat step 2 and 3 at a given time, where 100 times is the default value.
5. Return the results.

Data Simulation

The data is simulated in a way that clusters are clearly separable and visible, with a few realistic issues such as noise and outliers. The model generates five-dimensional data and consists of the following subsets:

- Subset 1 (cluster, 150 points): normal distribution with mean vector (1, 0, 1), and covariance matrix $0.1I_3$.
- Subset 2 (cluster, 175 points): normal distribution with mean vector (4, 3, 3) and a covariance matrix with diagonal element 0.5 and covariance 0.25 in all off-diagonals.
- Subset 3 (cluster, 175 points): normal distribution with mean vector (-3, -3, -2) and a covariance matrix with diagonal element 0.6 and covariance 0.4 in all off-diagonals.
- Subset 4 (noise, 10 points): uniform distribution on $[-3, 1]^3$.
- Subset 5 (noise, 10 points): uniform distribution on $[-1, 4]^3$.

All points existing multiple times in the bootstrap can only be used once in the bootstrap sample to avoid upweighted issues when computing the Jaccard mean (Hennig, 2007). The iteration chosen is 100 times as recommended by the function default. Besides, the resampling method used is bootstrap which is normally used to present ideas of variance and bias in the statistical (Hennig, 2007). It is noted that the cluster found by the clustering methods in the original data is likely to be the true cluster. Data is scaled before applying the cluster bootstrap function to avoid the dominance of a group feature.

Comparison between K-means and Hierarchical Clustering Methods

Table 4 - K-means

Clusterwise Jaccard mean:	0.9968	0.9973	0.9996
dissolved:	0	0	0
recovered:	100	100	100

K-means (K = 3): K-mean can be considered as nearly perfect for finding all clusters. Firstly, all clusters are recovered successfully at 100 times resampling. Besides, over 0.99 for Jaccard mean is recorded for all three clusters, with none time dissolved. Therefore, clusters found in the original dataset are highly possible to be true clusters.

Table 5 - Single linkage

Clusterwise Jaccard mean:	0.9832	0.8004	0.7706
dissolved:	0	21	23
recovered:	96	79	77

Hierarchical (K = 3, single linkage): Hierarchical method with single linkage performs well on detecting the first cluster as it indicates 96 times of clusters are recovered and 0.9832 for the Jaccard bootstrap mean. The second and third clusters are often found properly, shown by 0.8004 and 0.7706 for the Jaccard bootstrap mean, and 79 and 77 times recovered, respectively.

Table 6 - Complete linkage

Clusterwise Jaccard mean:	0.6749	0.3755	0.7937
dissolved:	5	81	0
recovered:	21	4	62

Hierarchical (K = 3, complete linkage): Hierarchical clustering method does not find any cluster reliably. Over 100 sampling times, the second cluster only recovered 4 times, but dissolved 81 times. Moreover, the Jaccard mean for this cluster is 0.3755, which implies that it should not be trusted. The third cluster has the highest Jaccard mean (0.7937), which indicates moderate stability of the cluster. Generally, single linkage performs better than complete linkage in the dataset. Although complete linkage is less sensitive to outliers and noise, it may breach the “closeness”, meaning that points can be closer to points in other clusters than its own cluster (Taylor, 2012). Besides, the small stability values may indicate the instabilities in clustering methods or clusters.

Table 7 - Average linkage

Clusterwise Jaccard mean:	0.9802	0.9549	0.9776
dissolved:	0	0	0
recovered:	100	100	100

Hierarchical (K = 3, average linkage): Hierarchical method with average linkage found all clusters reliable. Over 100 times of resampling, three clusters have perfectly recovered all times, and high Jaccard means (over 0.95 for each cluster).

Table 8 - Centroid linkage

Clusterwise Jaccard mean:	0.8276	0.8708	0.9723
dissolved:	24	1	0
recovered:	75	75	100

Hierarchical (K = 3, centroid linkage): The third cluster is considered as highly stable. It has successfully recovered over 100 times and has a relatively high Jaccard mean (0.9723). The first and the second cluster have both recovered 75 times with the Jaccard mean of 0.8276 and 0.8708, respectively. However, the stability of the first cluster should be considered as it is dissolved 24 times.

In general, after comparing all linkages' performance in terms of Jaccard mean, average linkage shows the best result with three highly stable clusters. Also, the centroid linkage is relatively good as it well finds two clusters. In contrast, the complete linkage shows the worst result because none of the clusters is found reliable. Compared to the hierarchical method, K-means shows a better result regarding cluster stability. It always finds the clusters perfectly. Besides, because the dataset was simulated to be clearly separated with only 3.8% noises, except hierarchical with complete linkage, the other algorithms found no pattern.

Because both K-means and hierarchical with average linkage show high stability of clusters (all over 0.99 and 0.95 respectively), it would be interesting to determine whether these clusters are the real ones. After detection, 5 points are clustered into the wrong groups by K-means, while only 1 point is wrongly allocated by the hierarchical method using average linkage. Although the hierarchical method using average linkage has a lower Jaccard mean than K-means on average over 100 times iteration, the clusters found by the hierarchical algorithm using average linkage perform better on finding the real clusters. This can reveal that K-means performance is more stable than hierarchical algorithms in this dataset. Therefore, notably, high stability may not necessarily indicate the validity of clusters.

According to Hennig (2007), the small stability values may refer to the meaningless clusters in the context of the true underlying model, or reveal the instabilities of clustering methods or clusters. Therefore, in this particular dataset, hierarchical with complete linkage is highly

possible to indicate unstable performance, while the performance of K-mean and hierarchical with average linkage can be considered as stable. However, it should be noted that the stability value may be affected by how bootstrap generates samples. Also, the structure of the dataset may support a clustering method over others (Thompson, 2019).