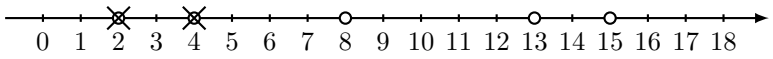
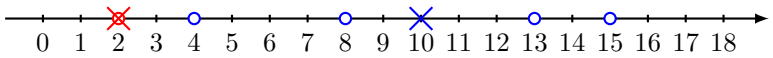
Aufgabe 1

- Aus den Datenpunkten werden zufällig k Clusterzentren ausgewählt.
- Jeder Datenpunkt wird zufällig einem der k Cluster zugeordnet.

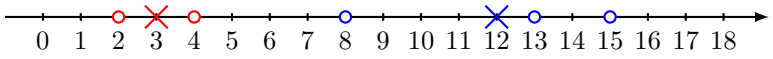
Aufgabe 2





$C_1 = 2/1 = 2 \qquad C_2 = (4 + 8 + 13 + 15)/4 = 10$

$J = 0^2 + 6^2 + 2^2 + 3^2 + 5^2 = 74$



$C_1 = (2 + 4)/2 = 3 \qquad C_2 = (8 + 13 + 15)/3 = 12$

$J = 1^2 + 1^2 + 4^2 + 1^2 + 3^2 = 28$

J: Summe der quadrierten Abstände aller Punkte von ihren Clusterzentren

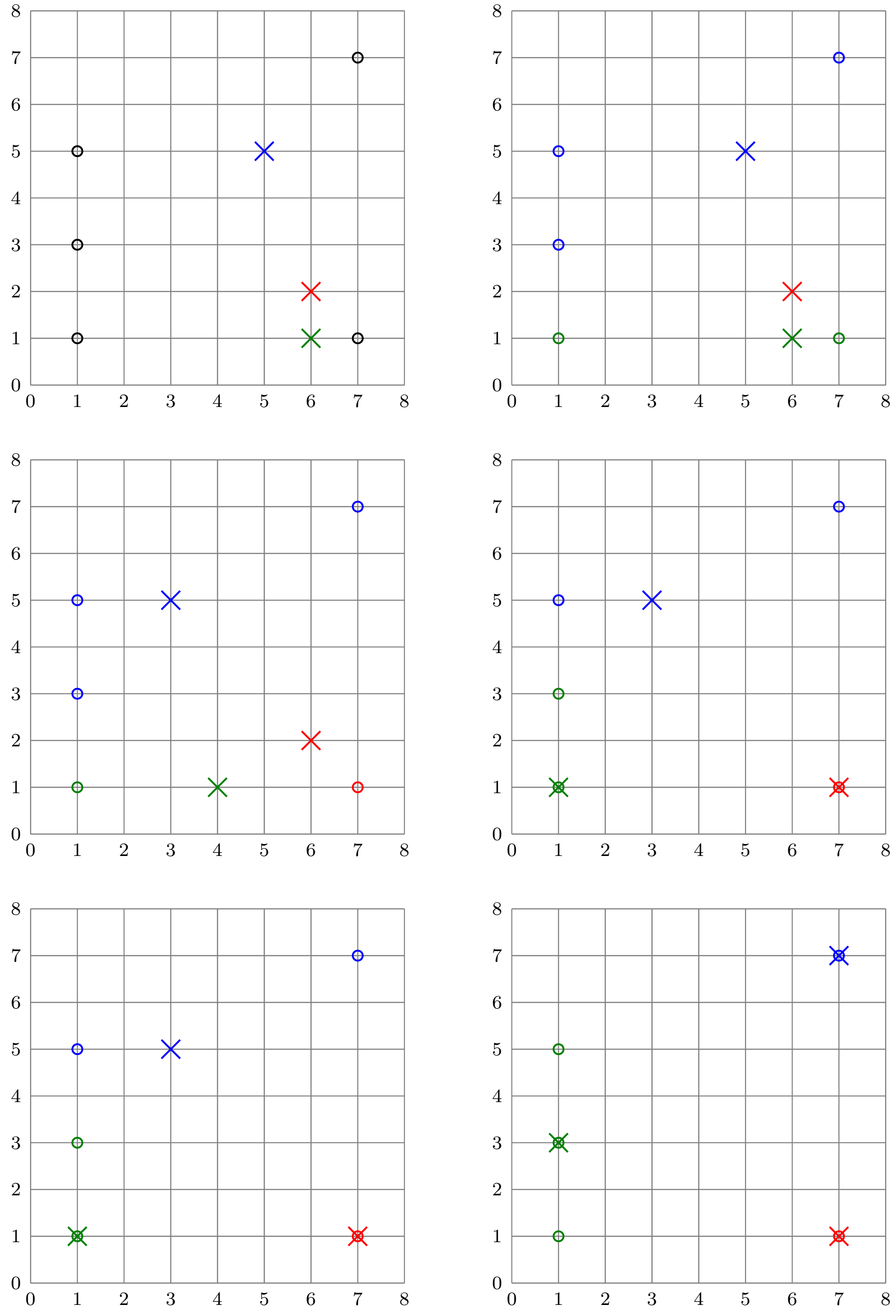
Aufgabe 3

Nein, der k-Means-Algorithmus findet im Allgemeinen ein *lokales Optimum*. Um das Ergebnis zu verbessern, kann man das Verfahren mehrfach mit jeweils anderen Startwerten durchführen, und dann das Resultat mit dem kleinsten Distortion Measure J wählen.

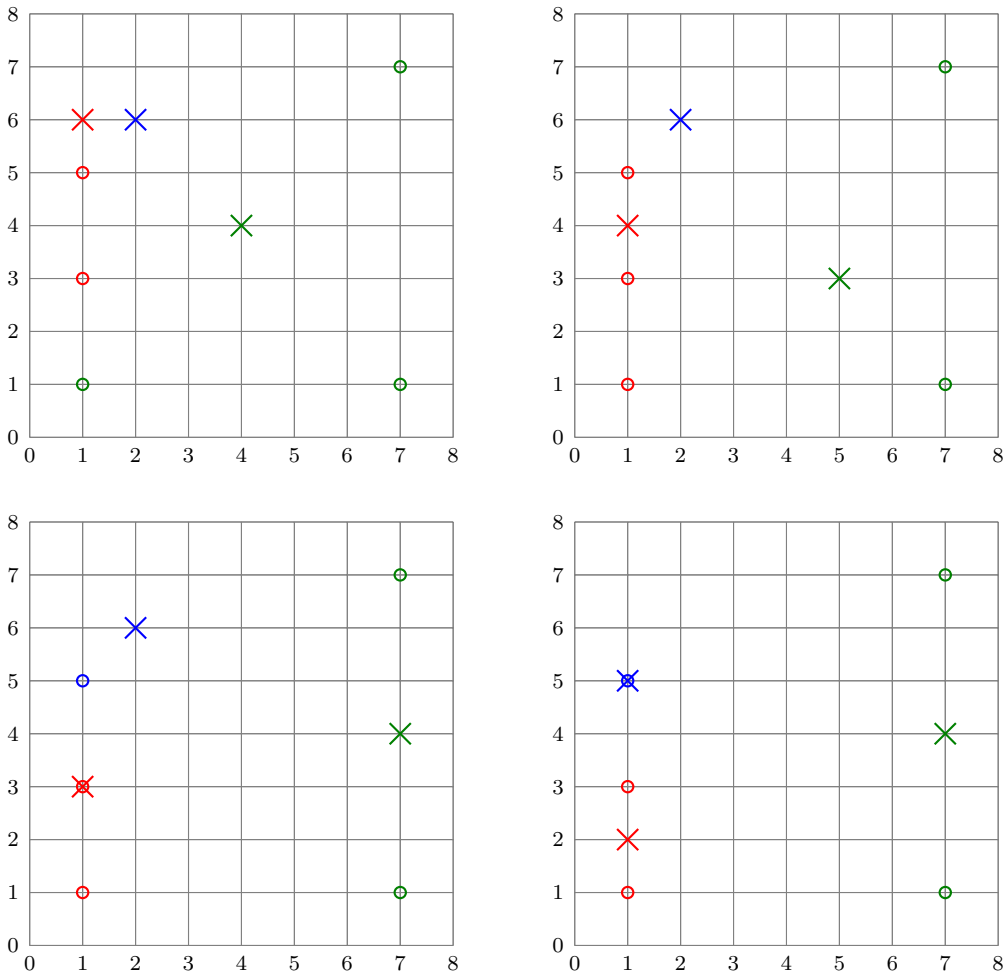
Aufgabe 4

- Markt- und Kundensegmentierung
- Bildsegmentierung
- Documente clustern
- Empfehlungssysteme

Aufgabe 5



Aufgabe 6



Aufgabe 7

Die Güte des Clusterings kann mit der Summe der Varianzen der einzelnen Cluster (*distortion measure*) gemessen werden.

$$J = \sum_{c_i \in \mathrm{C}} \sum_{p_j \in \mathrm{P}} |c_i - p_j|^2$$

Aufgabe 5: $J = 0^2 + 0^2 + 2^2 + 2^2 = 8$

Aufgabe 6: $J = 1^2 + 1^2 + 0^2 + 3^2 + 3^2 = 20$

Somit ist das Clustering in Aufgabe 5 besser.