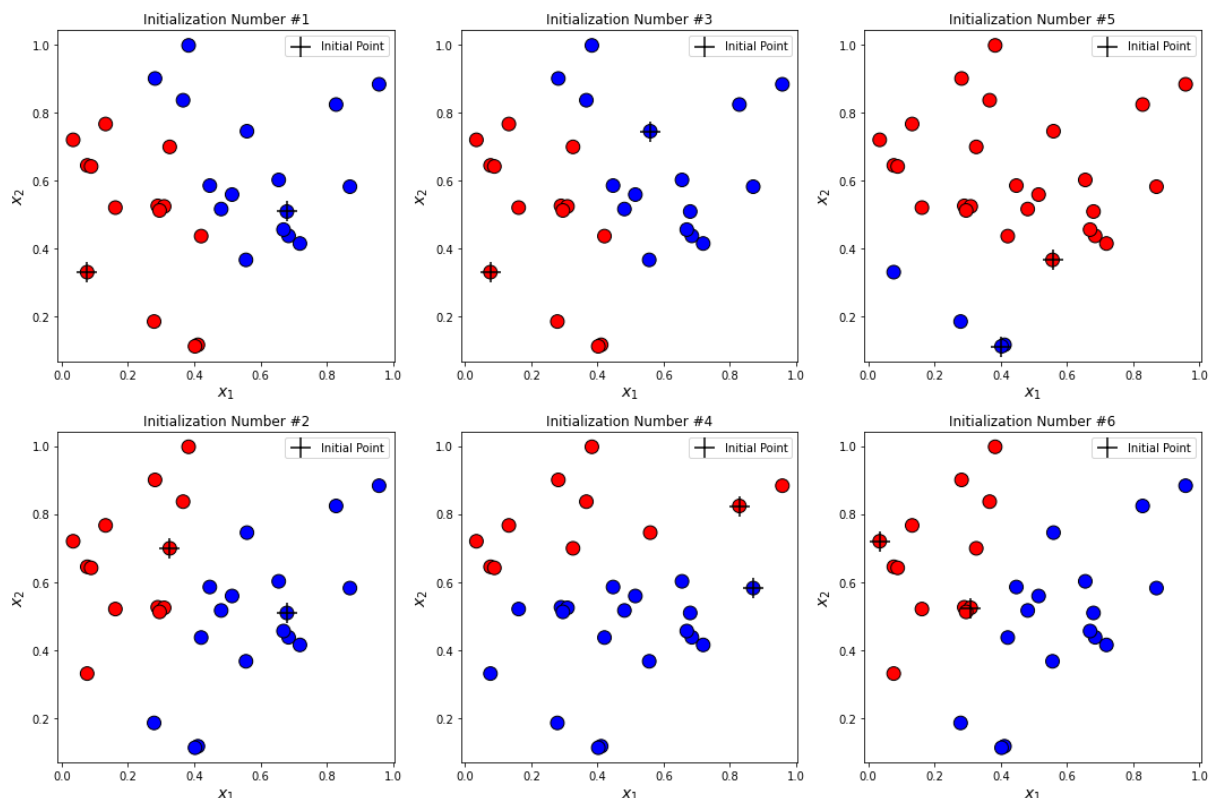


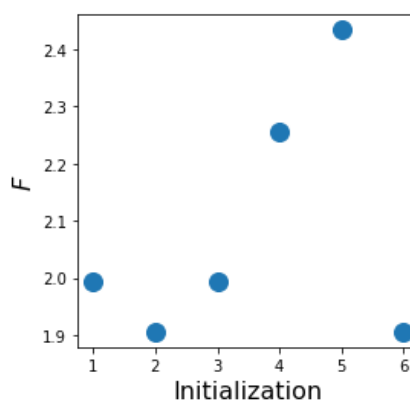
## Einfluss der Initialisierung beim $k$ -means Algorithmus

Das Ergebnis des Clusterings mit dem  $k$ -means Algorithmus ist abhängig von der Initialisierung, also der anfänglichen Wahl der Clusterzentren. Der Grund dafür ist, dass für eine festgelegte Anzahl an Clustern  $k$  in der Regel nur ein *lokales* Minimum der Kostenfunktion  $F$  gefunden wird.

Dies lässt sich an einem einfachen synthetischen zweidimensionalen Datensatz verdeutlichen. Im folgenden Beispiel wurde  $k = 2$  festgelegt und die beiden Clusterzentren (Initial Points) sechsmal zufällig initialisiert. Die Einfärbung der Datenpunkte zeigt die jeweils finalen vom Algorithmus gefundenen Cluster an.



Die gefundenen Lösungen unterscheiden sich zum Teil, abhängig von der Initialisierung, stark voneinander. Dies wird auch deutlich, wenn der Wert der Kostenfunktion  $F$  nach Erreichen des stationären Zustands für die unterschiedlichen Initialisierungen betrachtet wird.



$$F = \sum_{c=1}^k \sum_{\underline{x}^{(i)} \in D_c} \|\underline{x}^{(i)} - \underline{\mu}_c\|^2$$