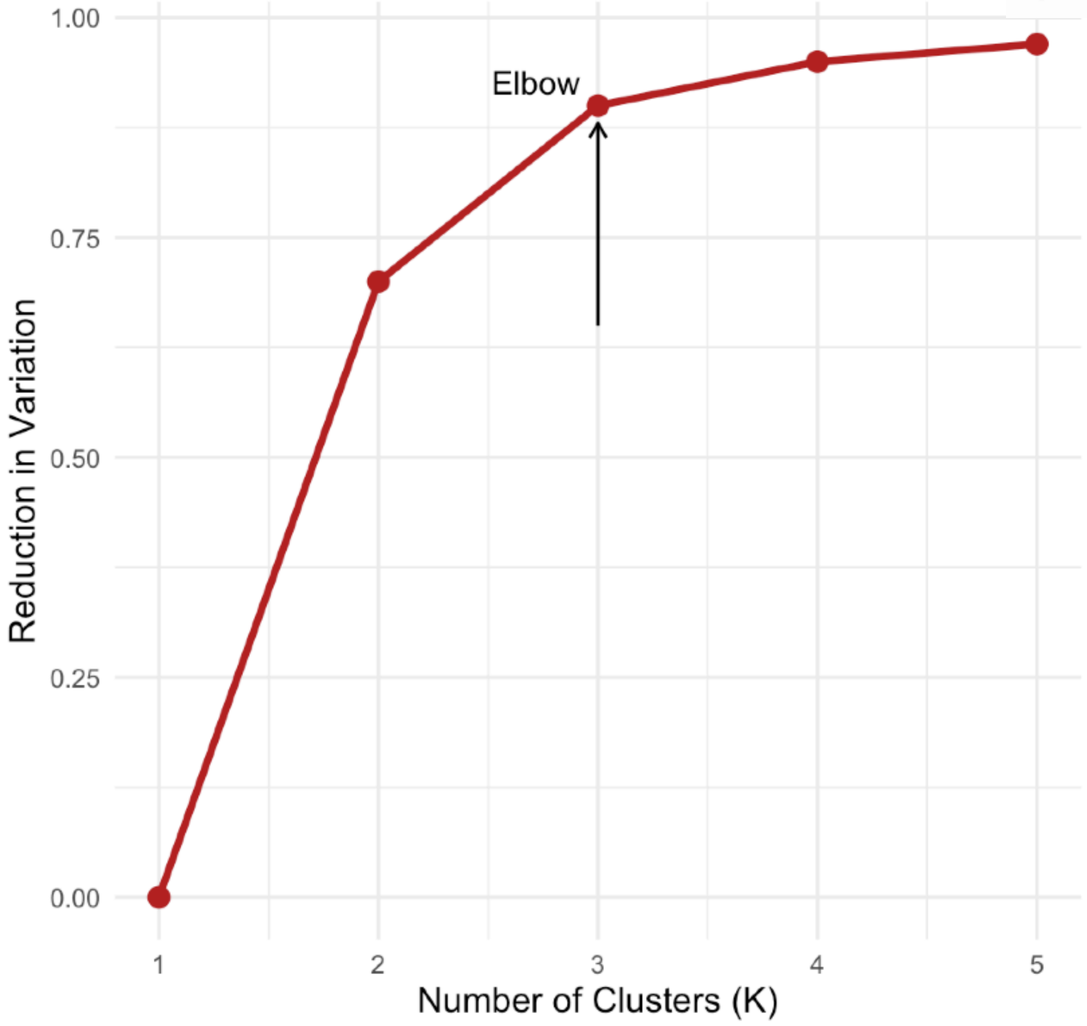
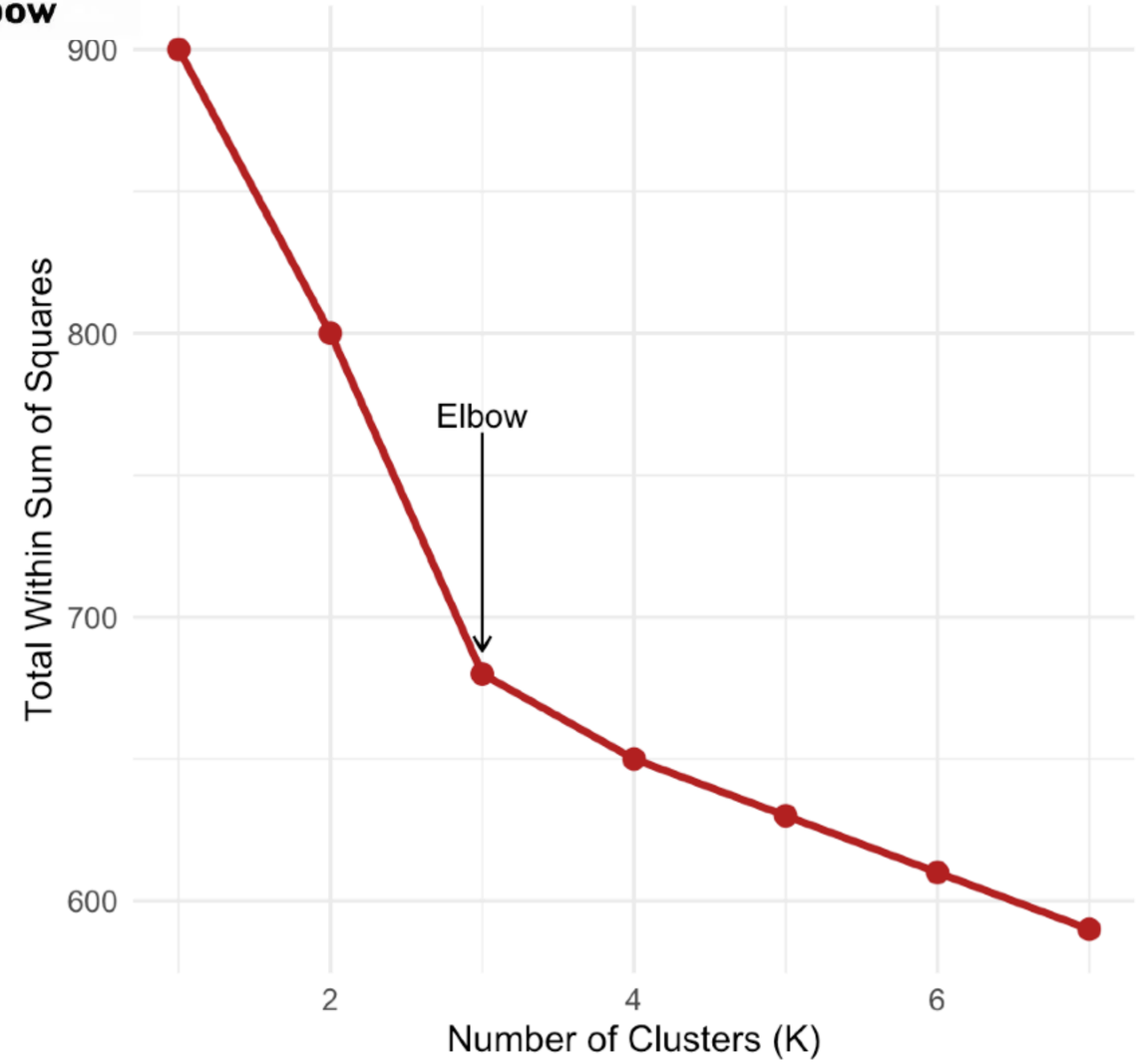
# How to decide on K

# Distortion

metric that assesses the performance of K-means (smaller values better)

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|x_n - \mu_k\|^2$$

actual data point $n$

center of cluster $k$

**Goal: choose $r_{nk}$ and $\mu_k$ that minimizes $J$**

**hard assignments!**

$$r_{nk} = \begin{cases} 1 & \text{if } k = \arg\min_j \|x_n - \mu_j\|^2 \\ 0 & \text{otherwise} \end{cases}$$

$$\frac{dJ}{d\mu_k} = 2 \sum_{n=1}^{N} r_{nk}(x_n - \mu_k) = 0 \implies \mu_k = \frac{\sum_n r_{nk} x_n}{\sum_n r_{nk}} = \frac{1}{N_k} \sum_n r_{nk} x_n$$

**optimal value for $\mu_k$ minimizing our loss is the mean of all data points in that cluster**