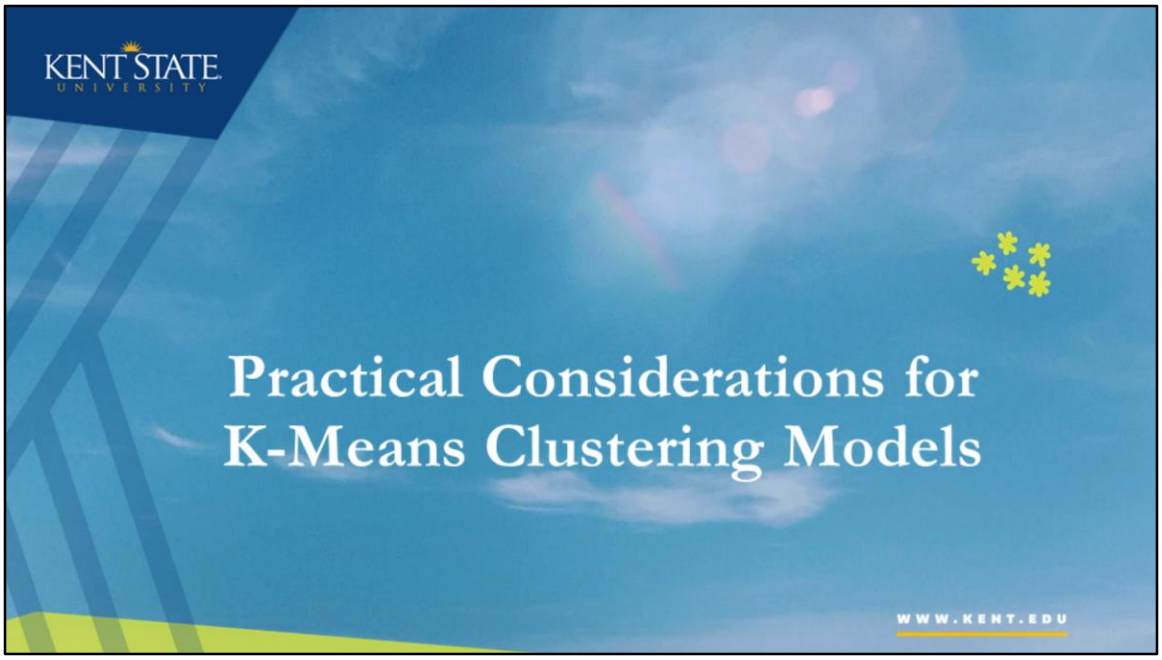
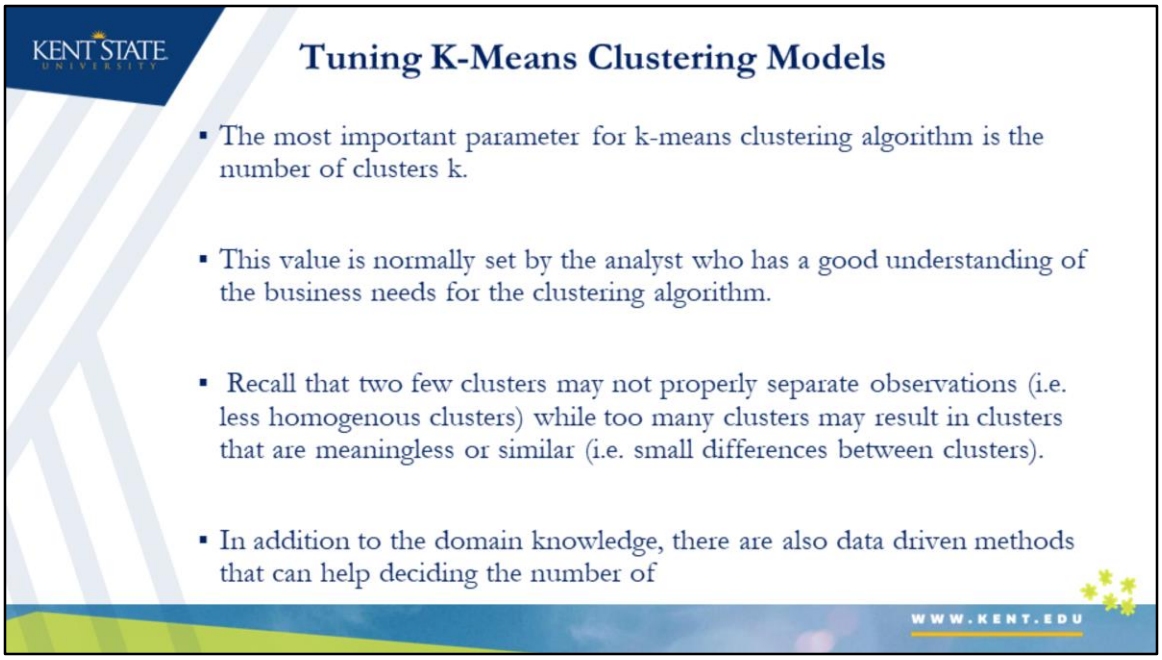
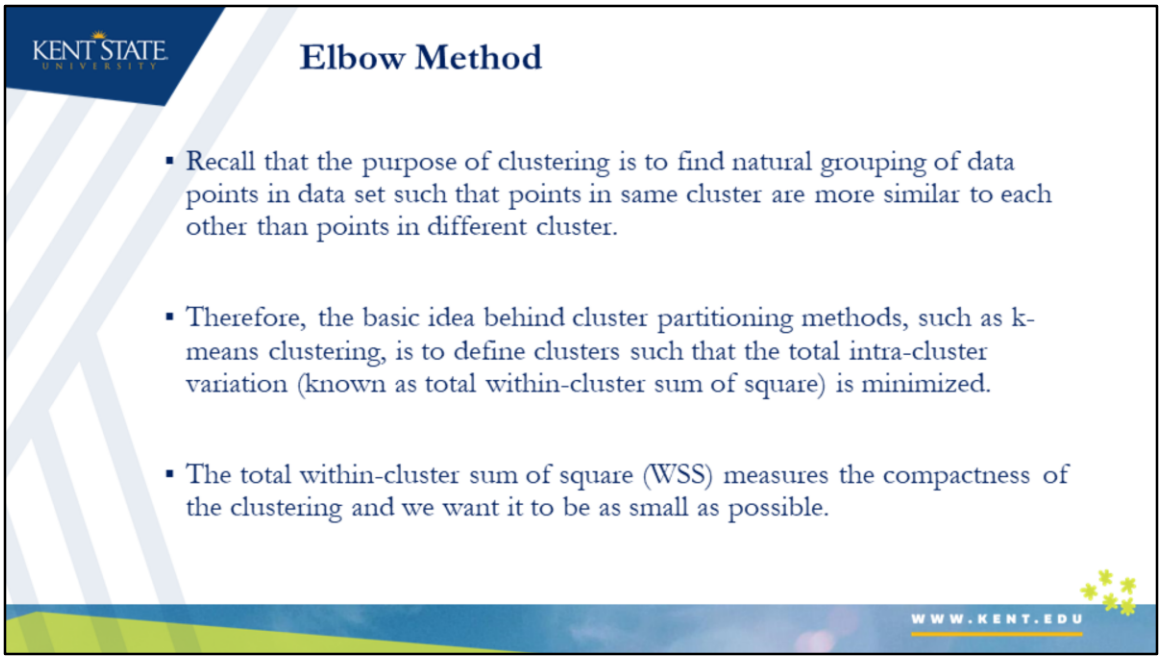


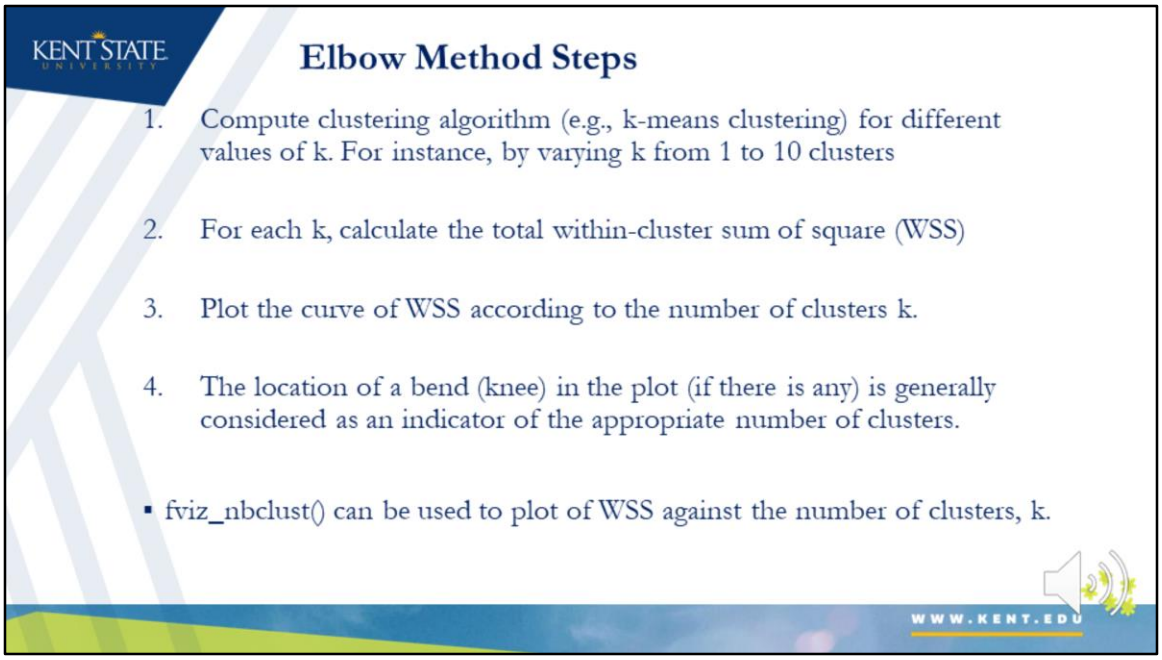
In this module, we look at some practical considerations for implementing k-means.



The most important parameter to set here is k, the number of clusters. The choice of the number of clusters can either be driven by external considerations (e.g., previous knowledge, practical constraints, etc.), or we can try a few different values for k and compare the resulting clusters. After choosing k, the n records are partitioned into these initial clusters. If there is external reasoning that suggests a certain partitioning, this information should be used. Alternatively, if there exists external information on the centroids of the k clusters, this can be used to initially allocate the records. Here, we will look at some data-driven methods to select k.



Recall the idea behind clustering. This is so that we group similar items into a cluster, with each cluster being different from other clusters. In other words, we are grouping items so that within cluster variation among items in a cluster is small compared to the variation between clusters. This within cluster variation is called the within-cluster sum of squares (WSS). The smaller WSS is, the less the difference between items in a cluster. This provides a metric to determine k.



An “elbow chart” is a line chart depicting the decline in cluster heterogeneity as we add more clusters. This simple chart allows us to graphically determine the best k for our data. Let us now apply this to our data.

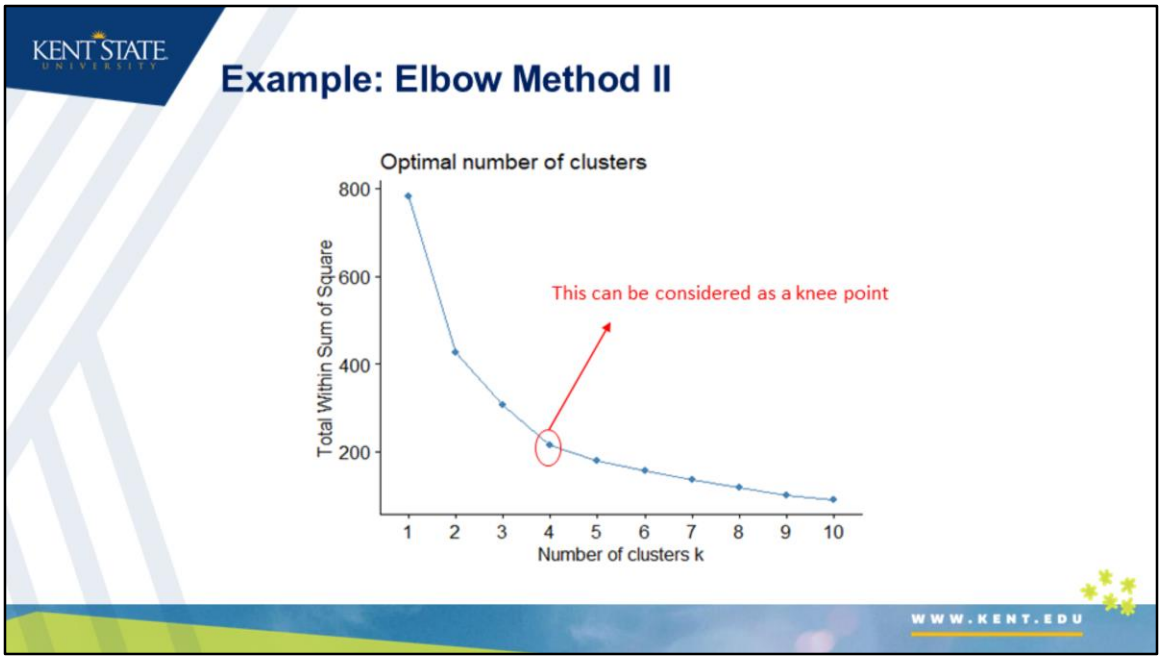
KENT STATE UNIVERSITY

Example: Elbow Method I

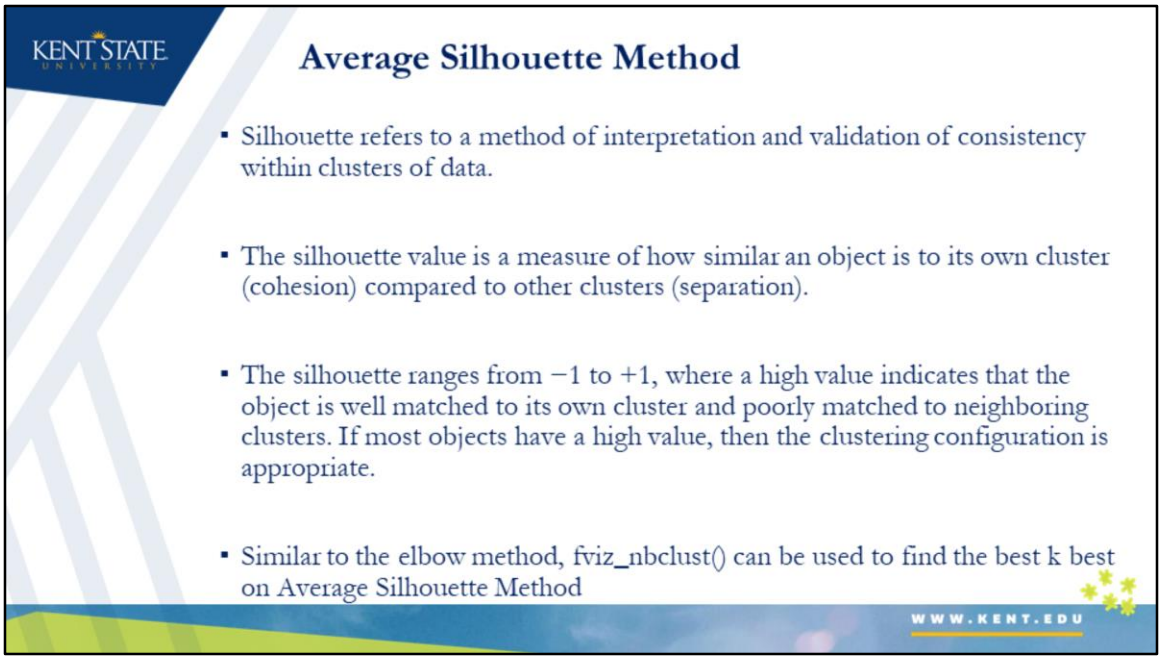
```
library(tidyverse)  # data manipulation
library(factoextra) # clustering & visualization
library(ISLR)
set.seed(123)

df<-Auto[,c(1,6)]
# Scaling the data frame (z-score)
df <- scale(df)
fviz_nbclust(df, kmeans, method = "wss")
```


This example shows how to implement the elbow method.



The chart shows that the elbow point 4 provides the best value for k. While WSS will continue to drop for larger values of k, we have to make the tradeoff between overfitting, i.e., a model fitting both noise and signal, to a model having bias. Here, the elbow point provides that compromise where WSS, while still decreasing beyond k = 4, decreases at a much smaller rate. In other words, k=4 provides the best value between bias and overfitting.

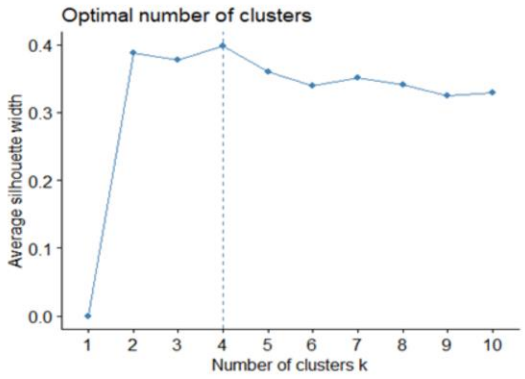


To ensure that the assignment of records to cluster is valid and consistent, we can apply the Silhouette Method. The idea is simple. The valid ranges for this method are between -1 and +1, and it measures how similar an object is to its assigned cluster compared to the other clusters. High values indicate a good matching of data to clusters.

KENT STATE UNIVERSITY

Example: Clustering of Cars

```
fviz_nbclust(df, kmeans, method = "silhouette")
```



WWW.KENT.EDU

Again, we see that 4 is the ideal number of clusters. Here we look for large values for the Silhouette Width (Y Axis). This concludes our discussion on k-means.