Class 7 Predictive Analytics for STP (I): Unsupervised Learning

Dr Wei Miao

UCL School of Management

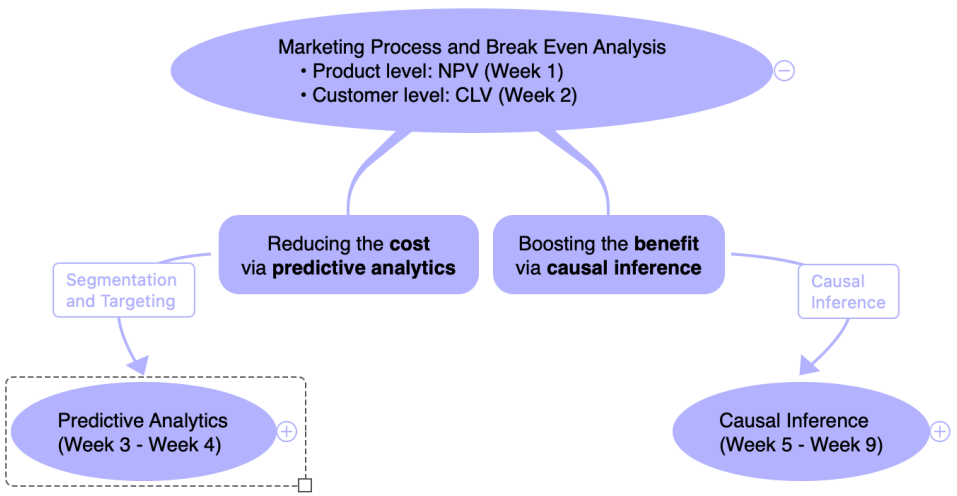
October 18, 2023

Section 1

Overview of Predictive Analytics

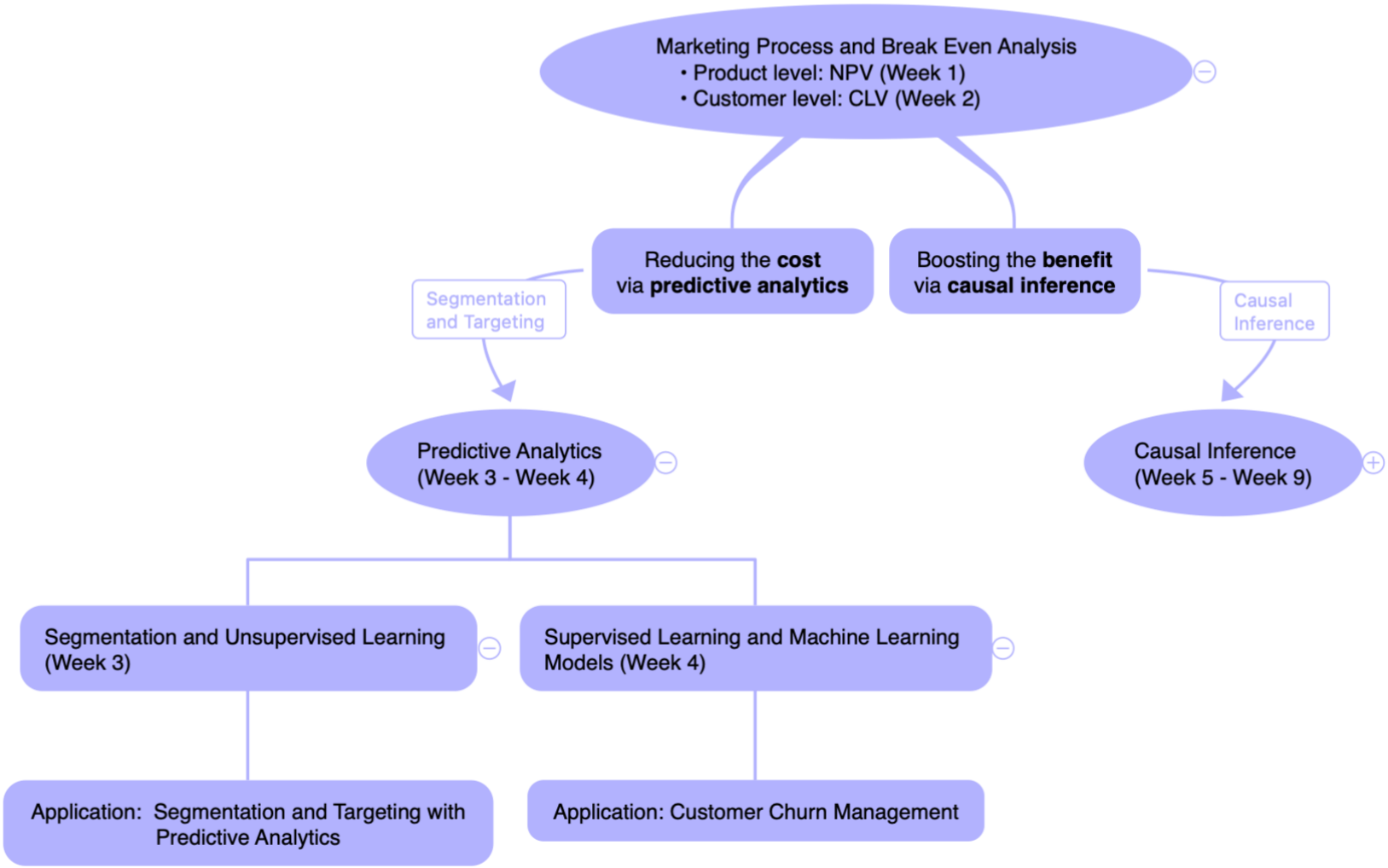
Our Journey So Far

- The core of any business decision is **break-even analysis** (cost-benefit analysis)
  - BEQ; NPV; CLV (Week 1 and Week 2)
- For better profitability management, we can work on either reducing the **cost** or boosting the **benefit**.



Roadmap of Predictive Analytics

- In Weeks 3 and 4, we will learn how to utilize predictive analytics to reduce marketing costs and improve marketing efficiency



Learning Objectives

- Understand the concept of statistical learning
- Understand the concept of unsupervised learning and how to apply clustering analyses for customer segmentation

Section 2

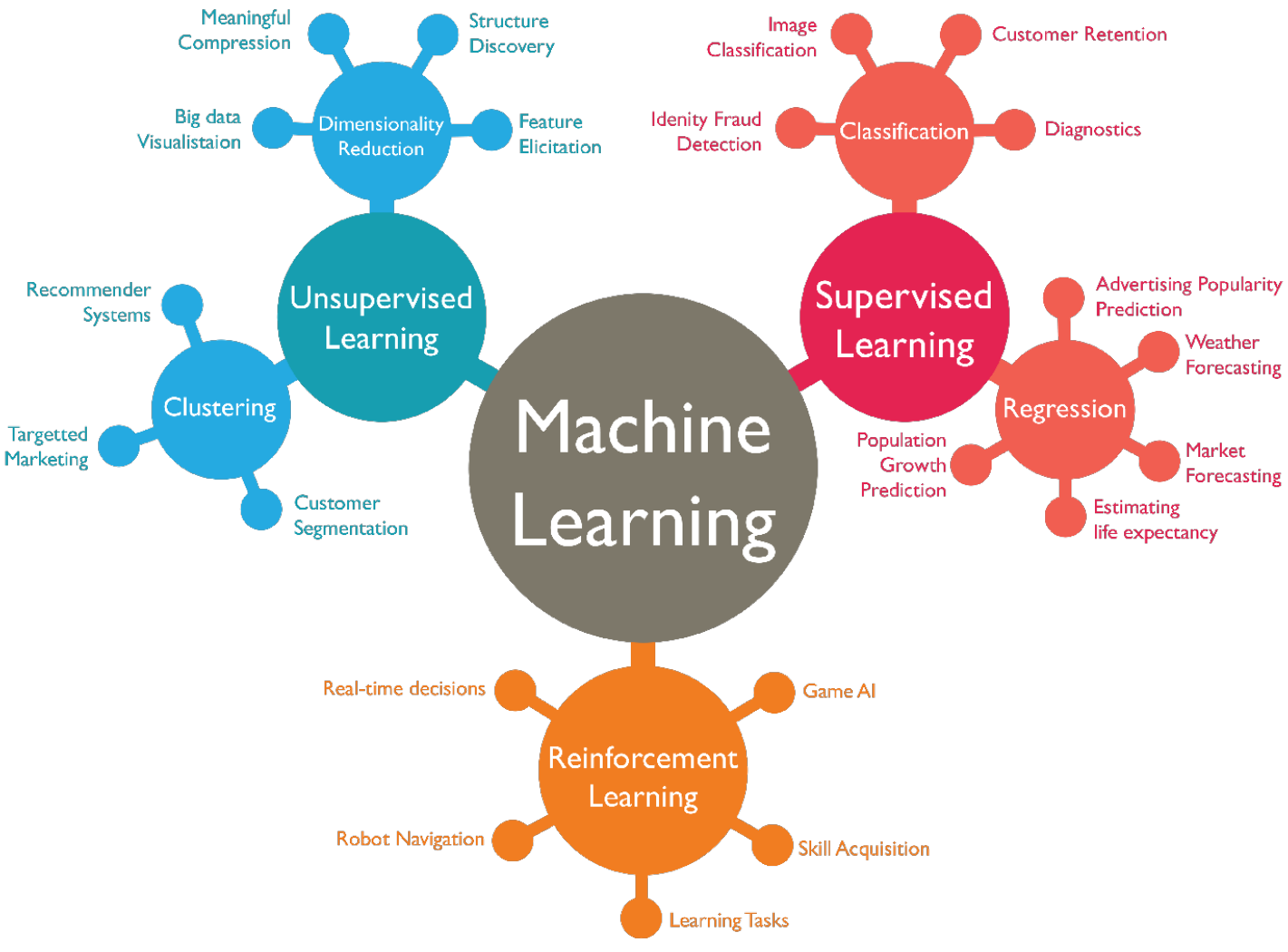
Predictive Analytics

Types of Predictive Analytics

- Unsupervised Learning
  - Only observe X => Want to uncover unknown subgroups
- Supervised Learning
  - Observe both X and Y => Want to predict Y for new data
- Reinforcement Learning
  - Rewards and punishments => Learn the best decision rules
  - Dynamic Coupon Targeting Using Batch Deep Reinforcement Learning: An Application to Livestream Shopping

In Term 2, you will learn predictive analytics models systematically. By then, think about how those techniques can be applied back to these case studies.

Types of Predictive Analytics



Section 3

Segmentation with Unsupervised Learning

Customer Segmentation

Segmentation is the process of dividing customers into meaningful groups based on any characteristics relevant to design and execution of your marketing strategy. It assumes that different customer groups offer different levels of value to the company and/or require different marketing programs to succeed with (e.g., based on different goals and needs).

Conventional Ways for Customer Segmentation

- **Customer value segmentation** is for targeting decisions based on customers' potential long-term financial and strategic value to your company.
- **Benefit segmentation** is for positioning and marketing mix design on the basis of customer and consumer goals or usage, the needs, wants, problems and the trade-offs they are willing to make across benefits (e.g., price vs. quality).
- **Psychographic segmentation** is for positioning and marketing mix design based on the psychology of the customer and consumer, including attitudes, identity, lifestyle, personality, etc.
- **Demographic segmentation** uses variables such as age, gender, income, family life cycle, educational qualification, socio-economic status, religion, company size and income, etc. These serve as proxies for goals, preferences or psychographics, as well as to characterize segments for marketing mix decisions.

Conventional segmentation methods require heavy human judgments. A more sensible way is to "let the data speak".

Commonly Used Clustering Algorithms

- K-means clustering
  - The number of clusters need to be pre-specified
- Hierarchical clustering
  - Observations are clustered in a tree-structured graph or dendrogram. No need to pre-determine the number of clusters.

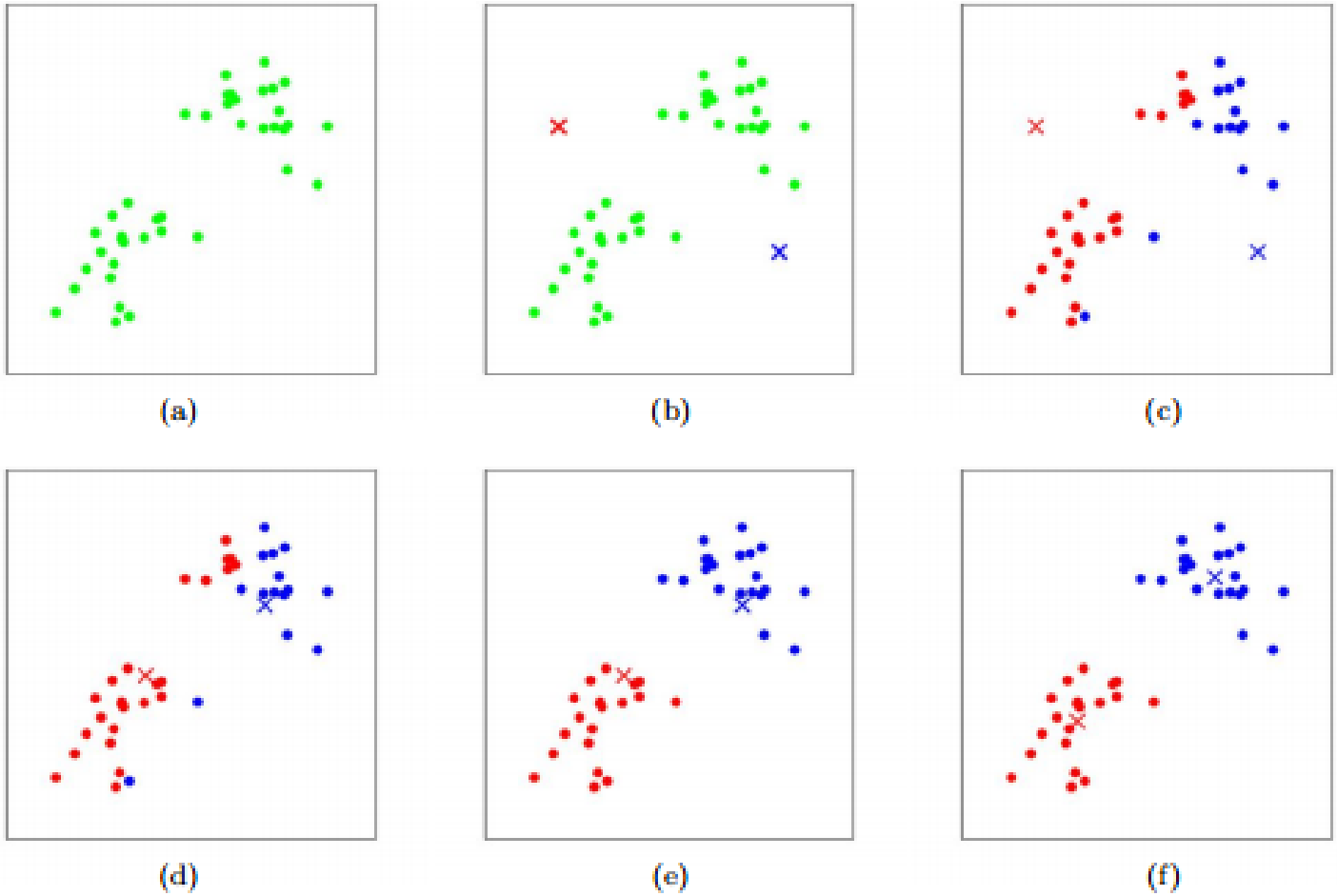
K-Means Clustering

K-means clustering is one of the most commonly used unsupervised machine learning algorithms for partitioning a given data set into a set of k groups (i.e. k clusters), where k represents the number of groups pre-specified by the analyst.

It can classify customers into multiple segments (i.e., clusters), such that customers within the same cluster are as similar as possible, whereas customers from different clusters are as dissimilar as possible.

- Input: customer data (characteristics of interest) and the number of clusters
- Output: clusters
  - Let $C_1, C_2, \cdots, C_k$ be the clusters
  - Every customer is categorized to only one of the clusters

K-Means Clustering: Intuition



Implementation of K-Means in R for Tesco

1. Decide to do customer segmentation based on *total spending* and *income*

```
kmeans(x, centers, iter.max = 10, nstart = 1,
       algorithm = c("Hartigan-Wong", "Lloyd", "Forgy",
                     "MacQueen"), trace=FALSE)
```

- `x`: data with selected variables to apply K-means
- `centers`: number of clusters
- `iter.max`: the maximum number of iterations allowed
- `nstart`: how many random sets should be chosen
- `algorithm`: which algorithm to choose; default often works
- `trace`: do you want to trace intermediate steps?

Implementation of K-Means in R for Tesco

- Need to re-scale the two variables using `scale()`, because the two variables are of very different scales
  - **This is extremely important!**
  - `set.seed()` is to allow replication of results. Refer to this data camp tutorial for more details.

```
set.seed(888)
data_kmeans <- data_full%>%
  select(Income,total_spending)%>%
  mutate(Income = scale(Income),
         total_spending = scale(total_spending))

result_kmeans <- kmeans(data_kmeans,
                        centers = 2,
                        nstart = 10)
```

Implementation of K-Means in R for Tesco

2. Examine the returned object, `result_kmeans`

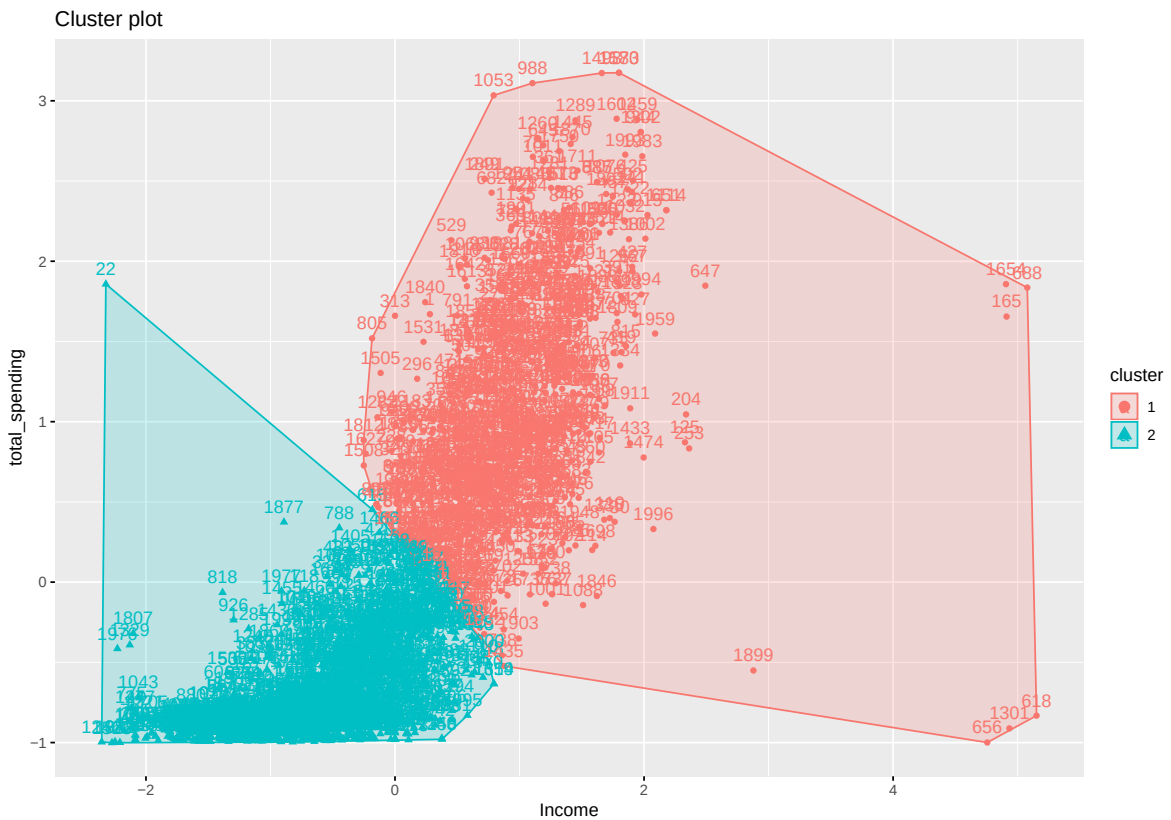
```
1 str(result_kmeans)
```

- `cluster`: A vector of integers (from 1:k) indicating the cluster to which each point is allocated.
- `centers`: A matrix of cluster centers.
- `totss`: The total sum of squares.
- `withinss`: Vector of within-cluster sum of squares, one component per cluster.
- `tot.withinss`: Total within-cluster sum of squares, i.e. sum(withinss).
- `betweenss`: The between-cluster sum of squares, i.e. $totss-tot.withinss$.
- `size`: The number of points in each cluster.

Implementation of K-Means in R for Tesco

3. Visualize the clusters
   - We need 2 packages `cluster` and `factoextra`
   - Use `fviz_cluster()` to generate visualizations

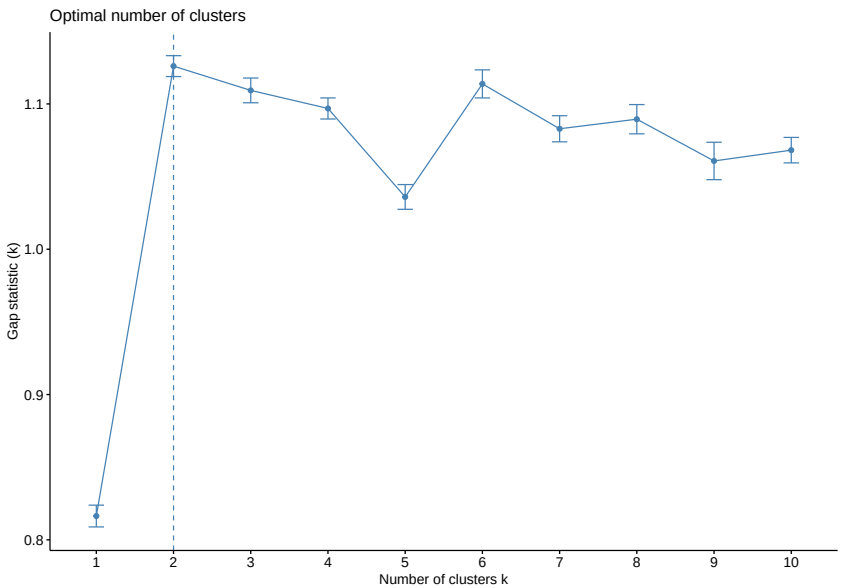
```
1 pacman::p_load(cluster,factoextra)
2 set.seed(888)
3 fviz_cluster(result_kmeans,
4               data = data_kmeans)
```



Implementation of K-Means in R for Tesco

4. Determine the optimal number of clusters using statistical criteria
   - **Gap** Method

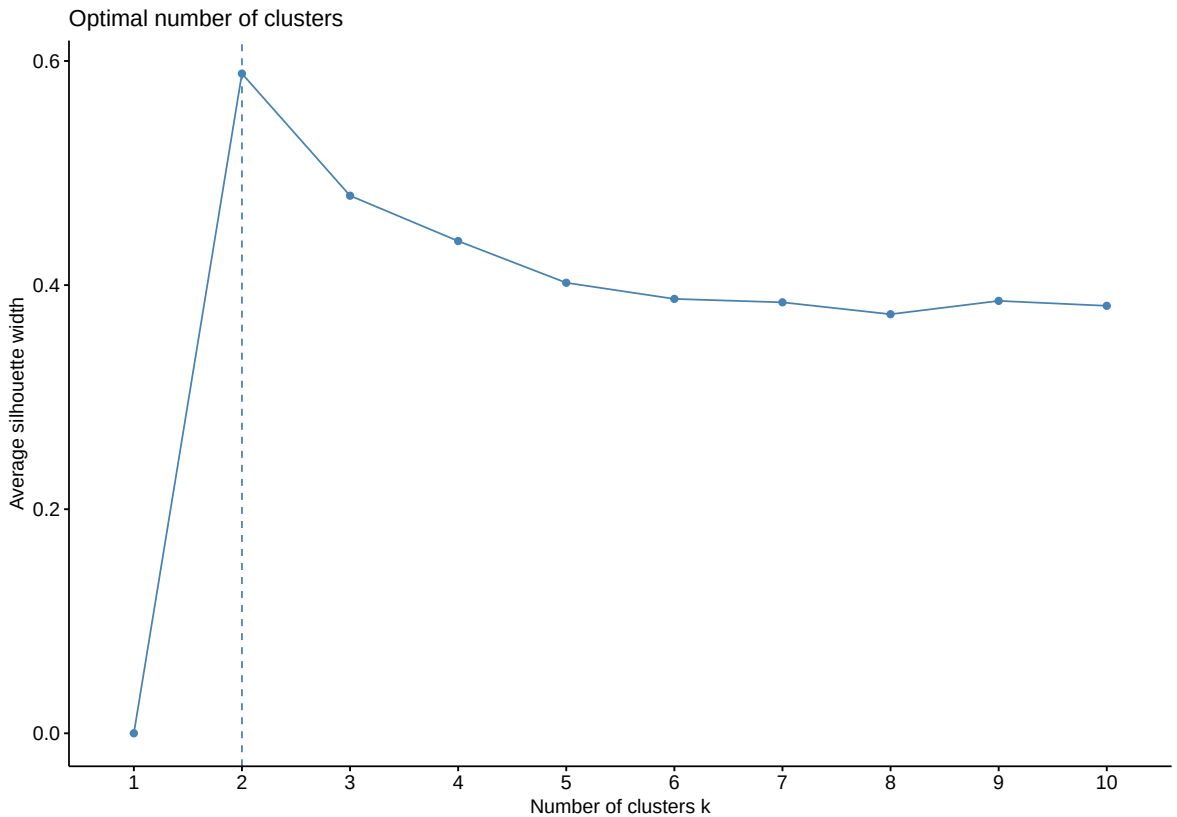
```
set.seed(888)
gap_stat <- clusGap(data_kmeans,
                    FUN = kmeans,
                    K.max = 10,
                    B = 50)
fviz_gap_stat(gap_stat)
```



Implementation of K-Means in R for Tesco

4. Determine the optimal number of clusters using statistical criteria
   - **Silhouette** method

```
set.seed(888)
fviz_nbclust(data_kmeans, kmeans, method = "silhouette")
```



Implementation of K-Means in R for Tesco

5. Compare the CLV in the two segments, and decide which segment to serve.
   - This is a general idea of segmentation and targeting using unsupervised learning
   - Finish this exercise after class

Pros and Cons of K-means Clustering

Advantages

- Easy to implement and explain
- Computationally efficient

Drawbacks

- As the number of variable increases, curse of dimensionality problem occurs
- Sensitive to outliers and initial seeds

After-Class Readings

- Useful source: K-means Cluster Analysis