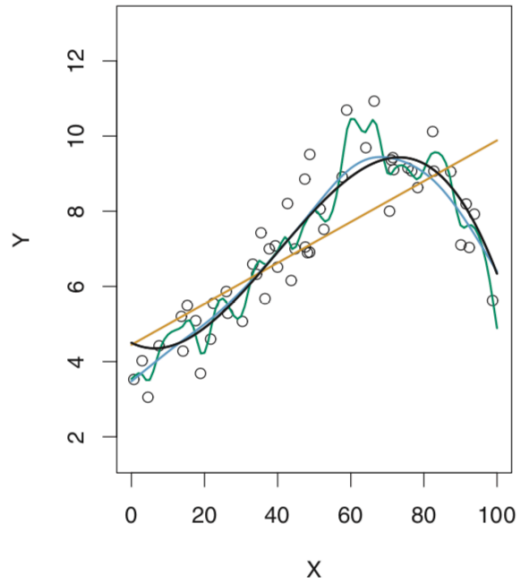
Unsupervised Learning

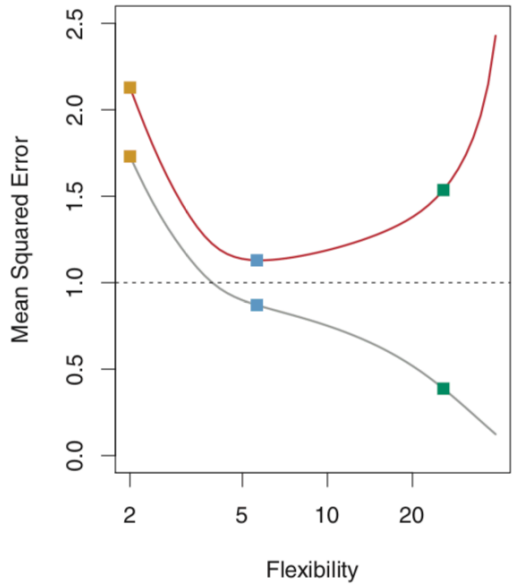
Clustering

Mathematics 699
6/5/2019
Yuta Hayashi
IHRTLUHC

Machine Learning

- Supervised Learning vs Unsupervised Learning
- Training and test error
- Variance-bias tradeoff
- Cross-validation





K-means Clustering

Objective: $$\underset{C_1,\ldots,C_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} W(C_k) \right\}$$

Where $$W(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2,$$

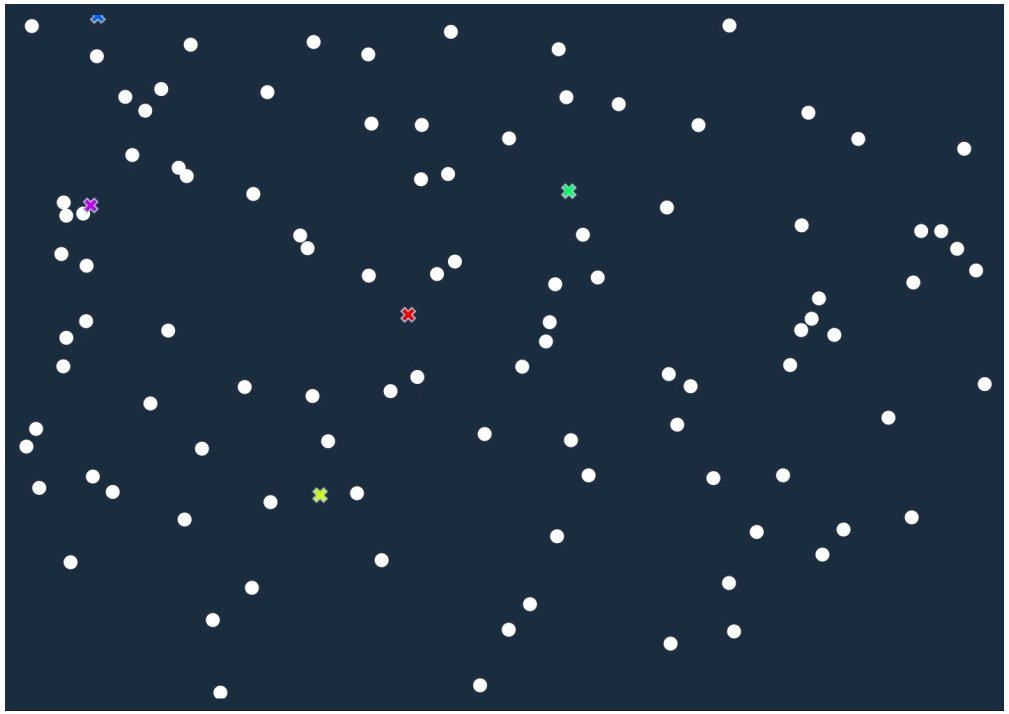
K-means Clustering

- Algorithm

  Initial Step: Randomly assign K centroids

  I. Assign each observation to the cluster whose centroid is closest
  II. For each cluster, compute the cluster centroids

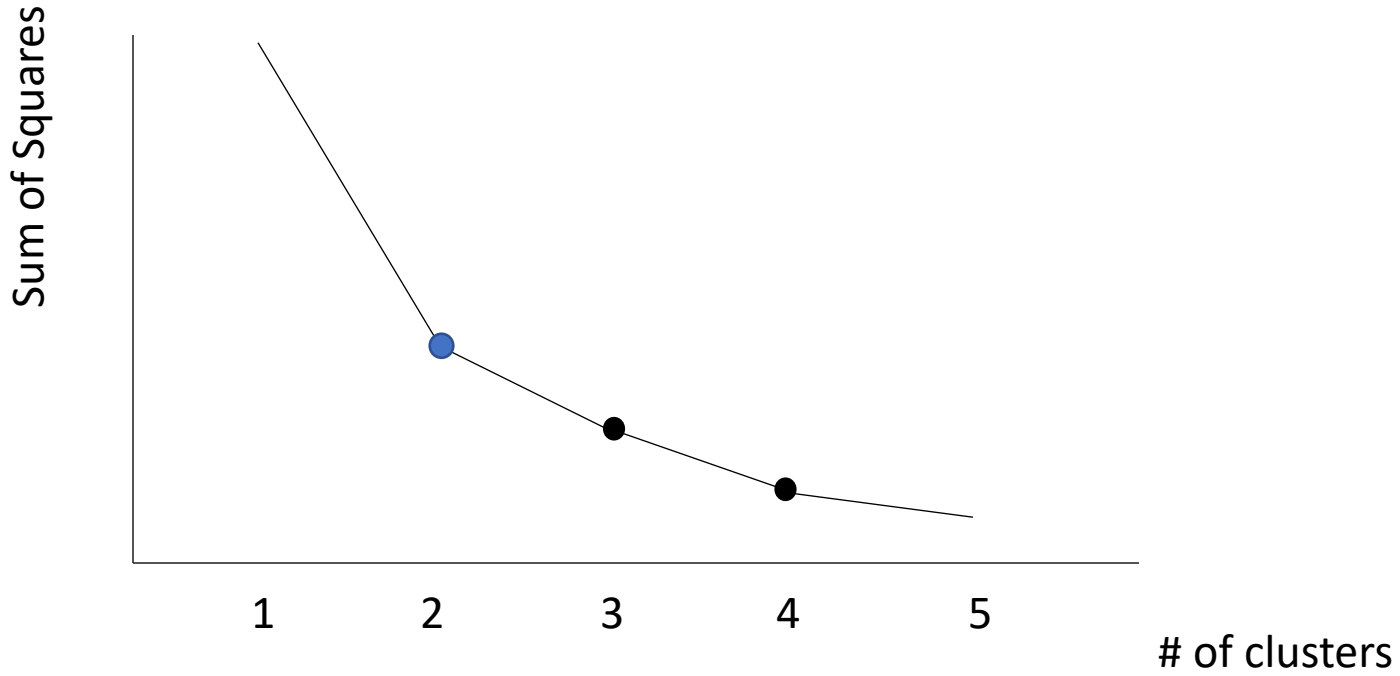
Iterate I and II until the cluster assignment stays the same



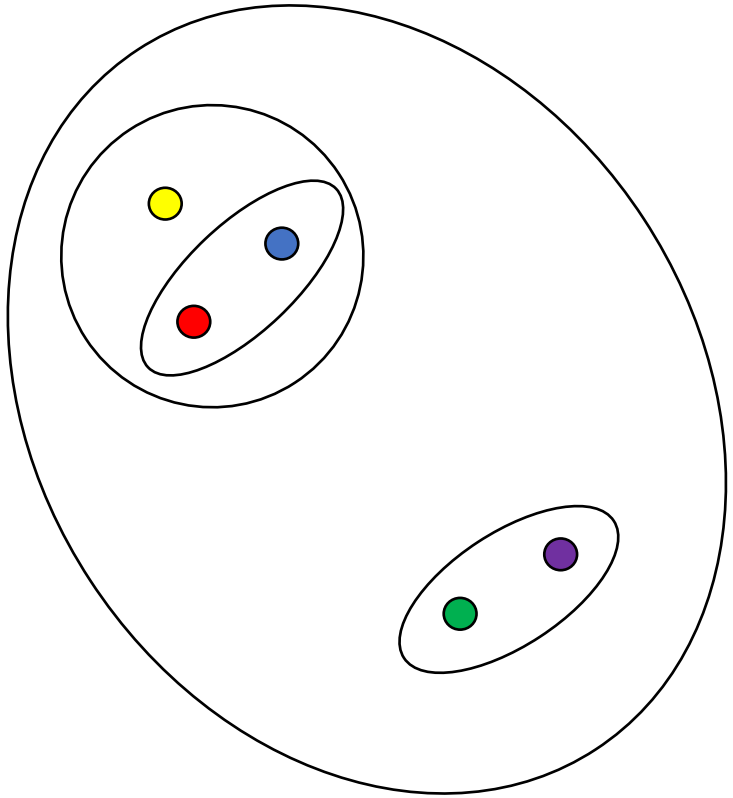
Cited from
http://tech.nitoyon.com/ja/blog/2013/11/07/k-means/

K-means Clustering

- Elbow Method
  - Scree plot:



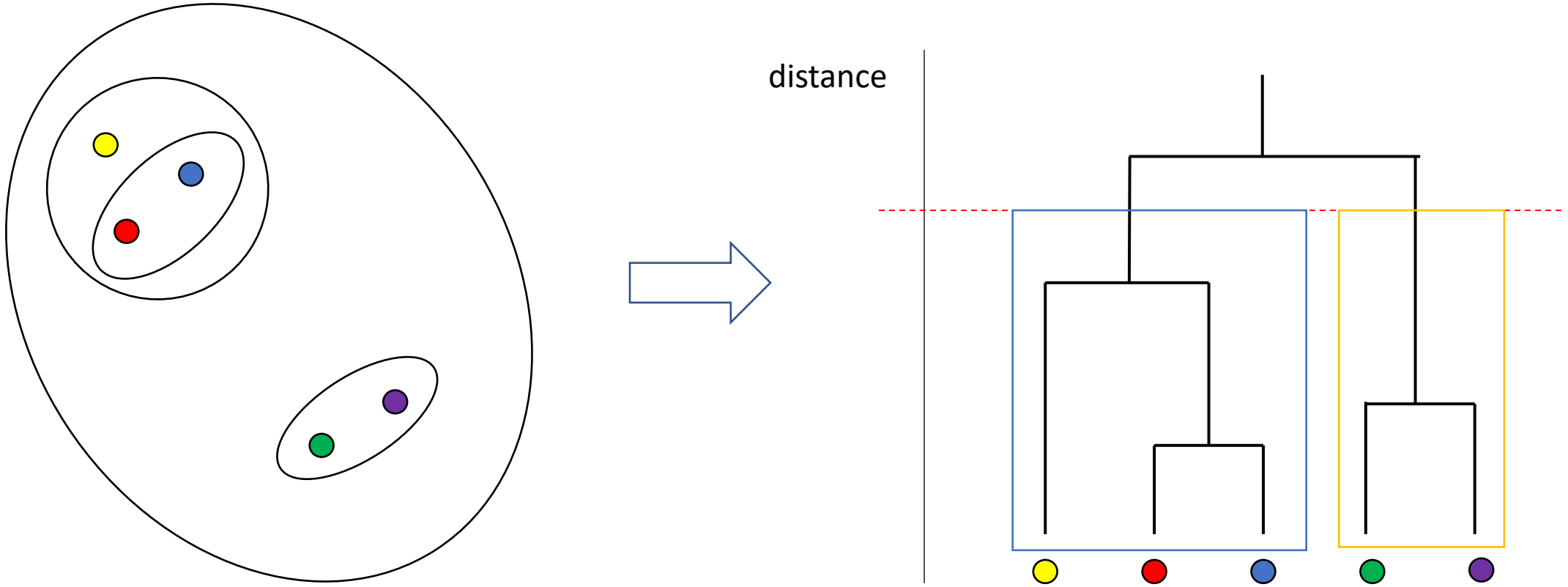
Hierarchical Clustering



5 clusters 4 clusters 3 clusters 2 clusters 1 cluster (each point)

Hierarchical Clustering

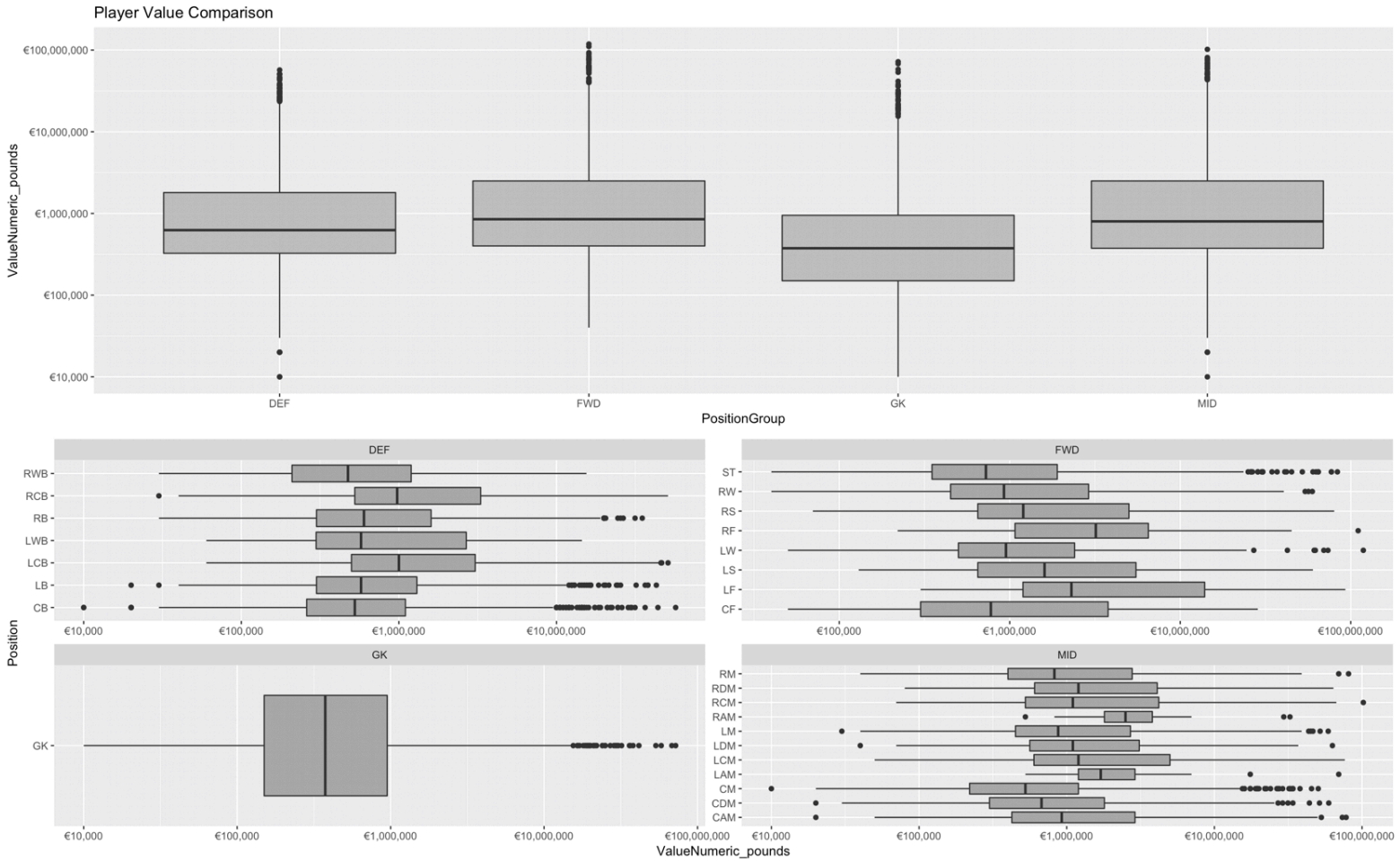
- Dendrogram



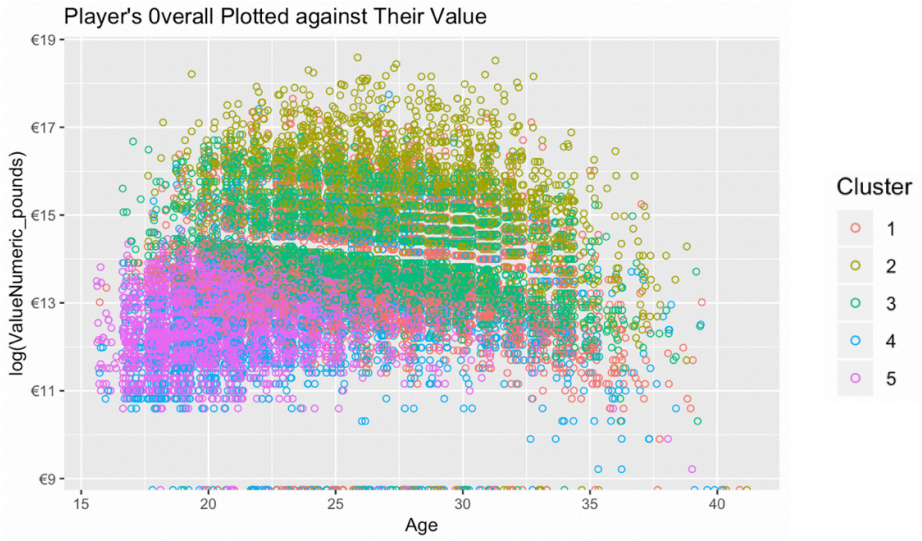
FIFA data

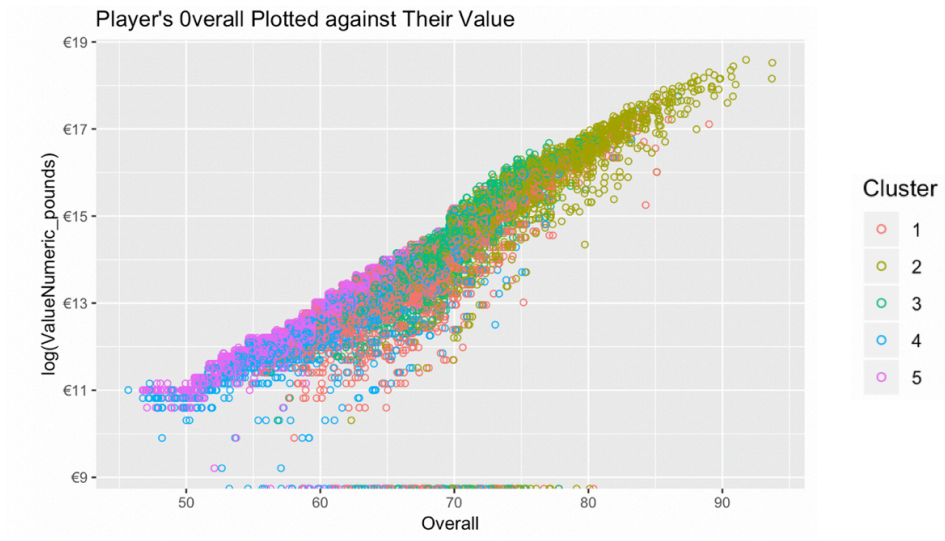
- Collected from Kaggle database, FIFA 2018 season all over the world
- 18207 rows(Players) and 95 columns(Traits)
  - 47 categorical 48 quantitative variable

Quick View


Player Value Comparison
ValueNumeric_pounds
€100,000,000
€10,000,000
€1,000,000
€100,000
€10,000
DEF
FWD
GK
MID
PositionGroup
DEF
RWB
RCB
RB
LWB
LCB
LB
CB
FWD
ST
RW
RS
RF
LW
LS
LF
CF
GK
GK
MID
RM
RDM
RCM
RAM
LM
LDM
LCM
LAM
CM
CDM
CAM
Position
ValueNumeric_pounds

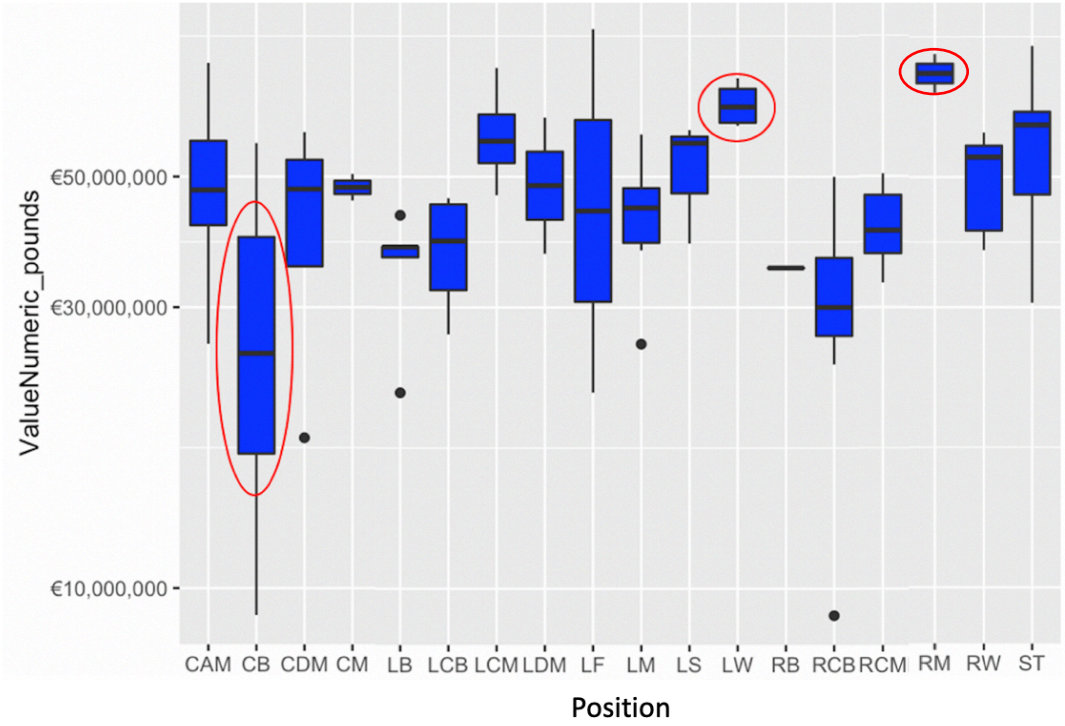
Clustering Analysis





	Defender	Forward	Midfielder
Cluster 1	140	1351	1471
Cluster 2	499	284	1490
Cluster 3	2675	57	2047
Cluster 4	2537	8	204
Cluster 5	15	1718	1626

Over/Under Compensation get rid of small samples



Cited from http://logfact.com/football-soccer-field-player-positions-abbreviations/

Conclusion

- Clustering is heavily based on compensation, overall ability, and positions
- If you want to be the BEST player in the world, DON’T be a CB player