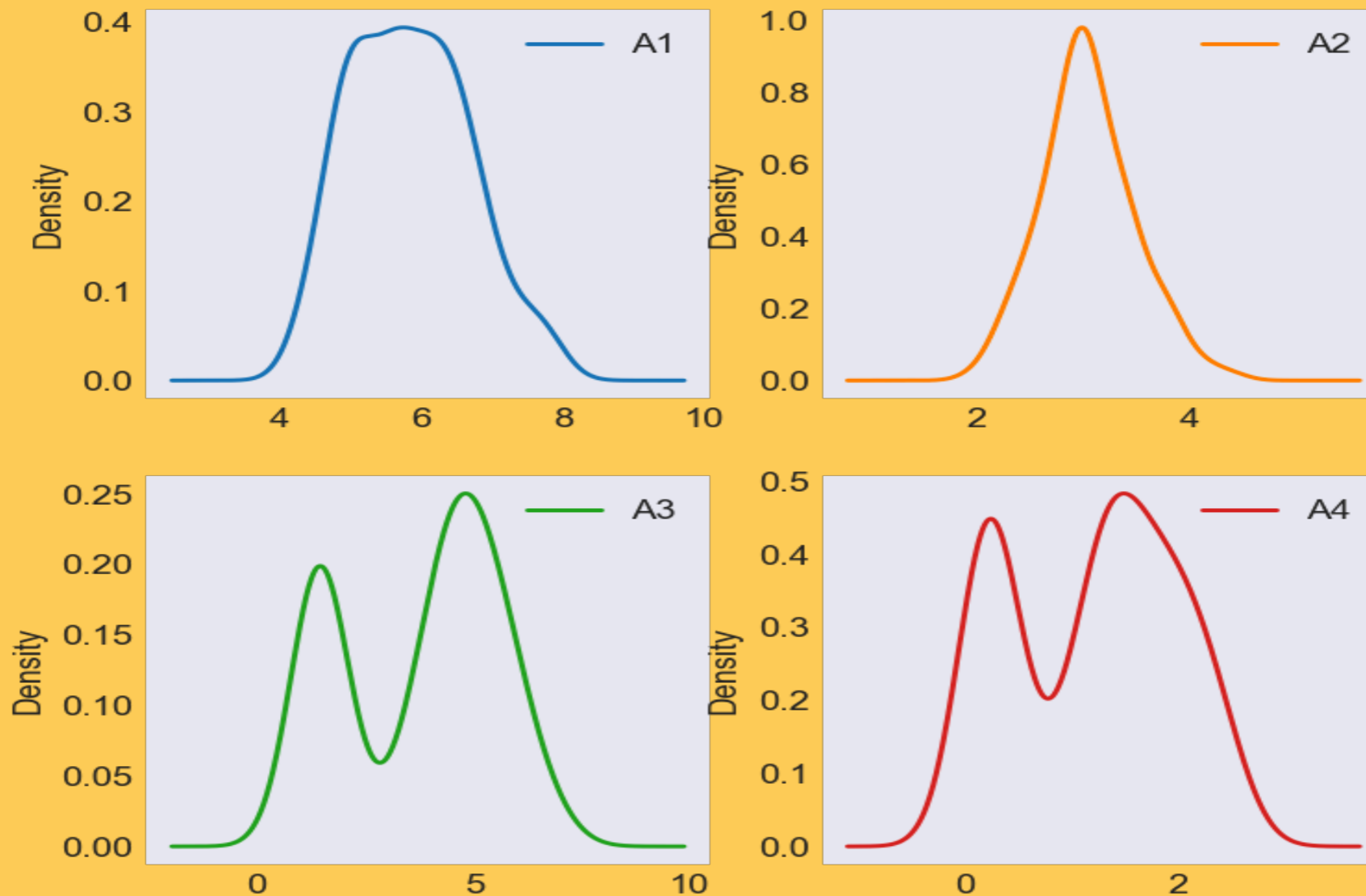ABHISHEK DAS

BCGDV MACHINE LEARNING ASSIGNMENT

SECTION 1

STANDARDIZATION VIA MIN–MAX SCALER



SCALES ARE OF SIMILAR ORDER BUT DISTRIBUTIONS DIFFERENT

K-MEANS CLUSTERING

▸ Intuitive: Randomly pick centroids and move them until they are 'center' of data clusters

▸ Distance: Minimize within cluster variance = Euclidean^2

▸ Cohesion: How far are points from their centroid?
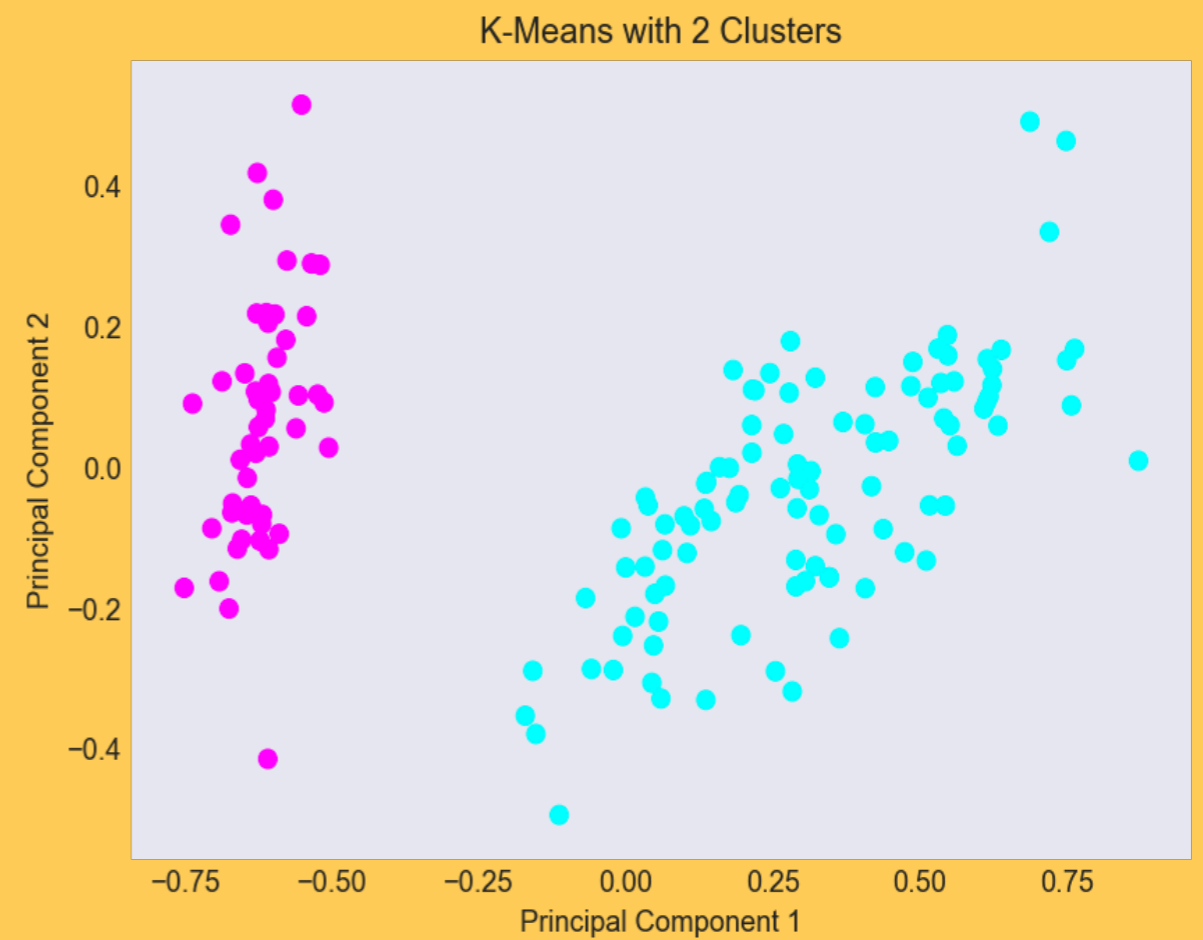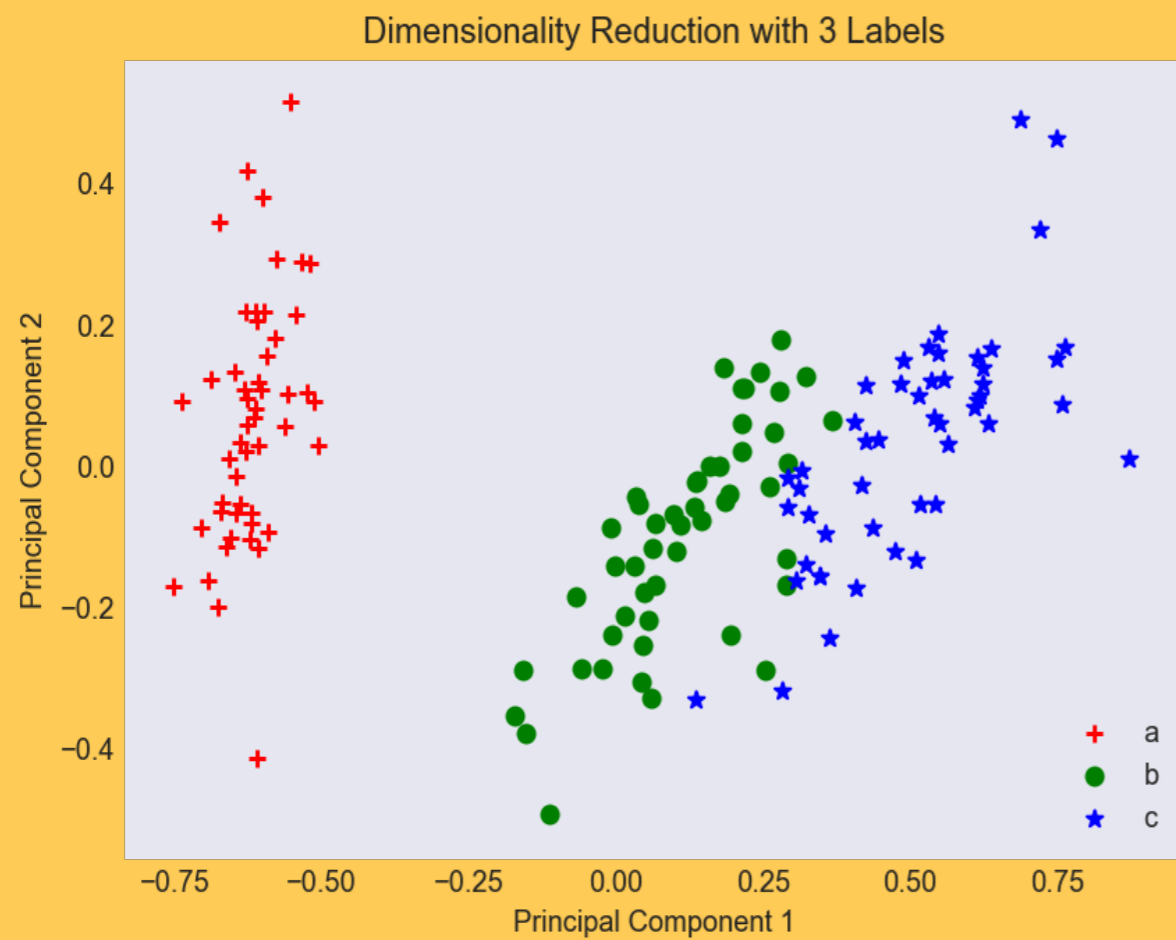
▸ Separation: How far are clusters from each other?
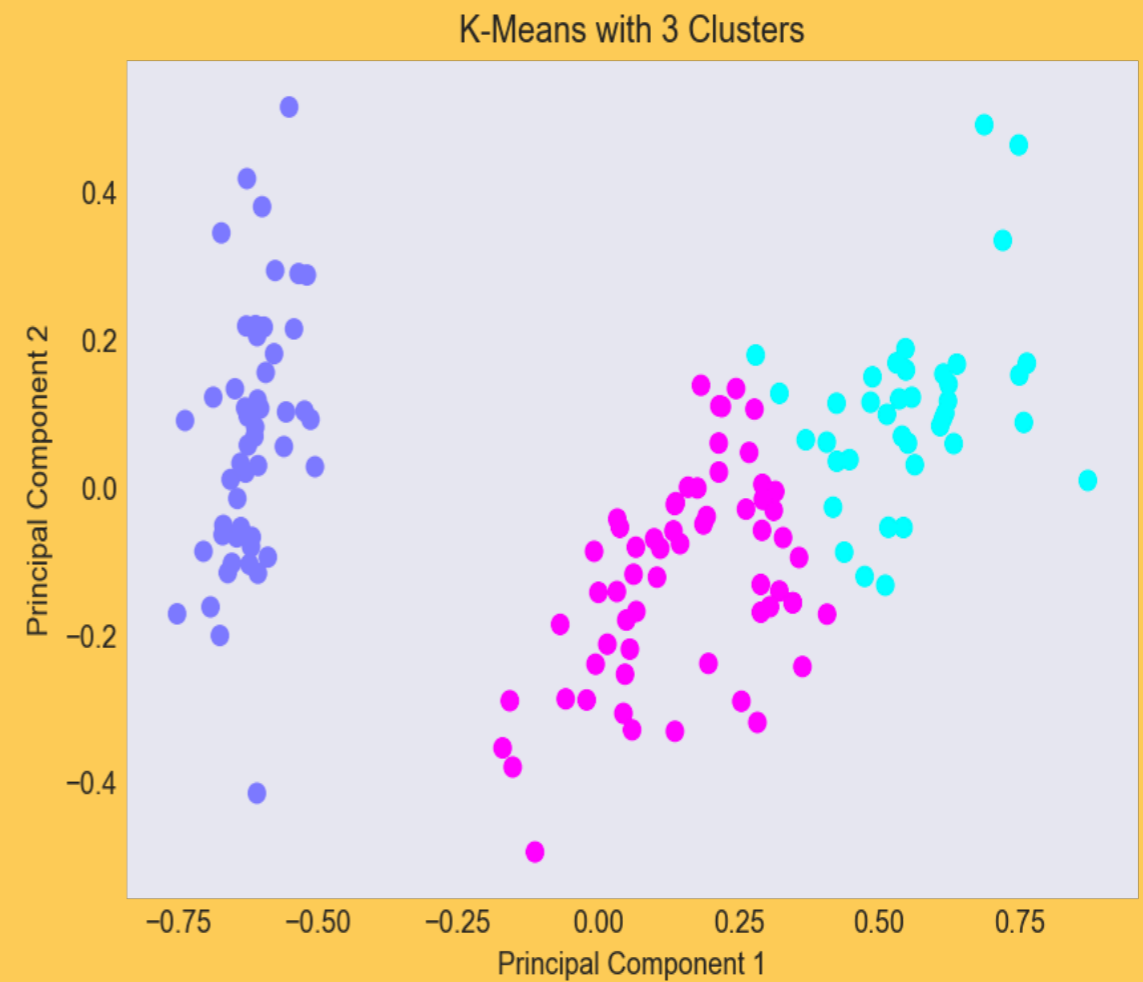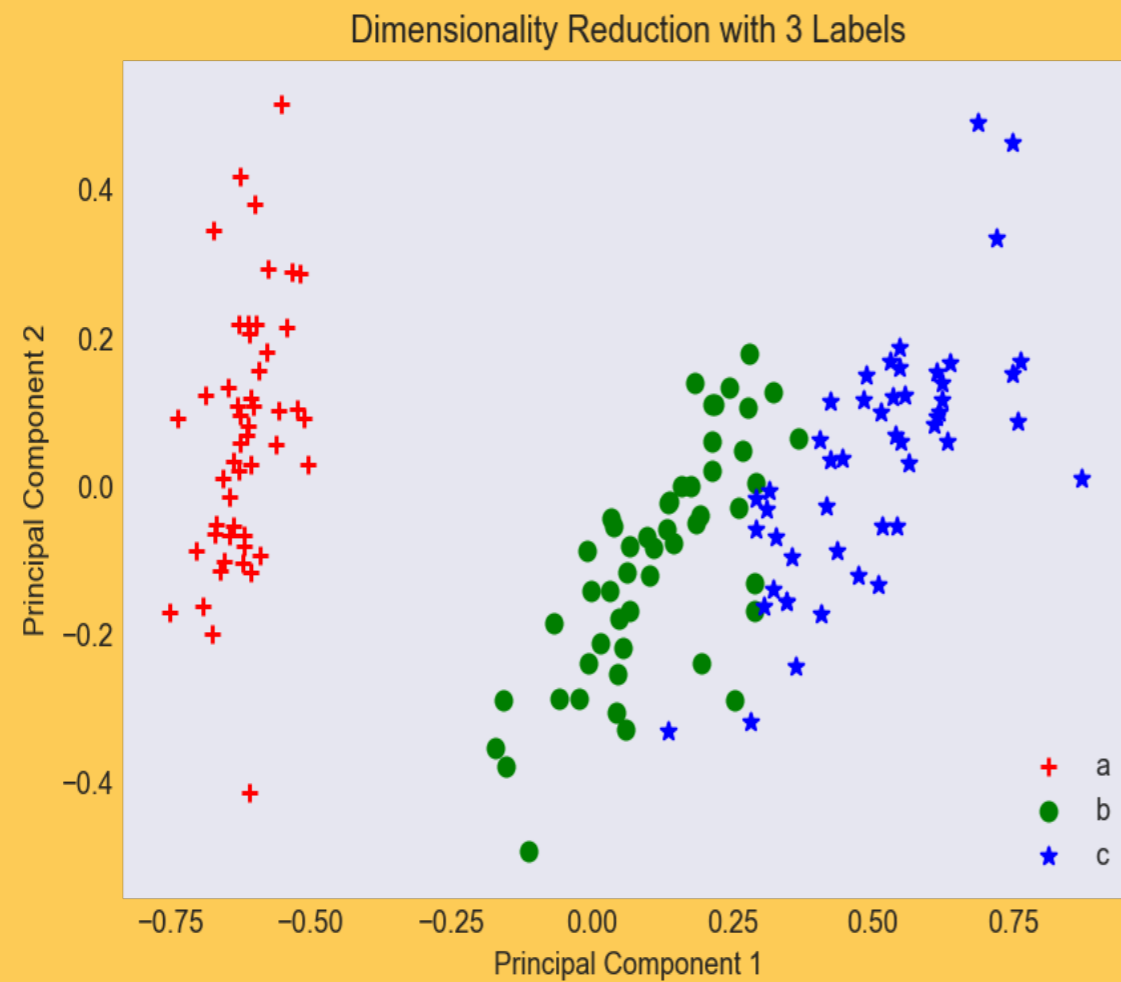
CHOOSING THE K IN K-MEANS

STANDARDIZED DATA

Highest Silhouette Coefficient with 2 Clusters
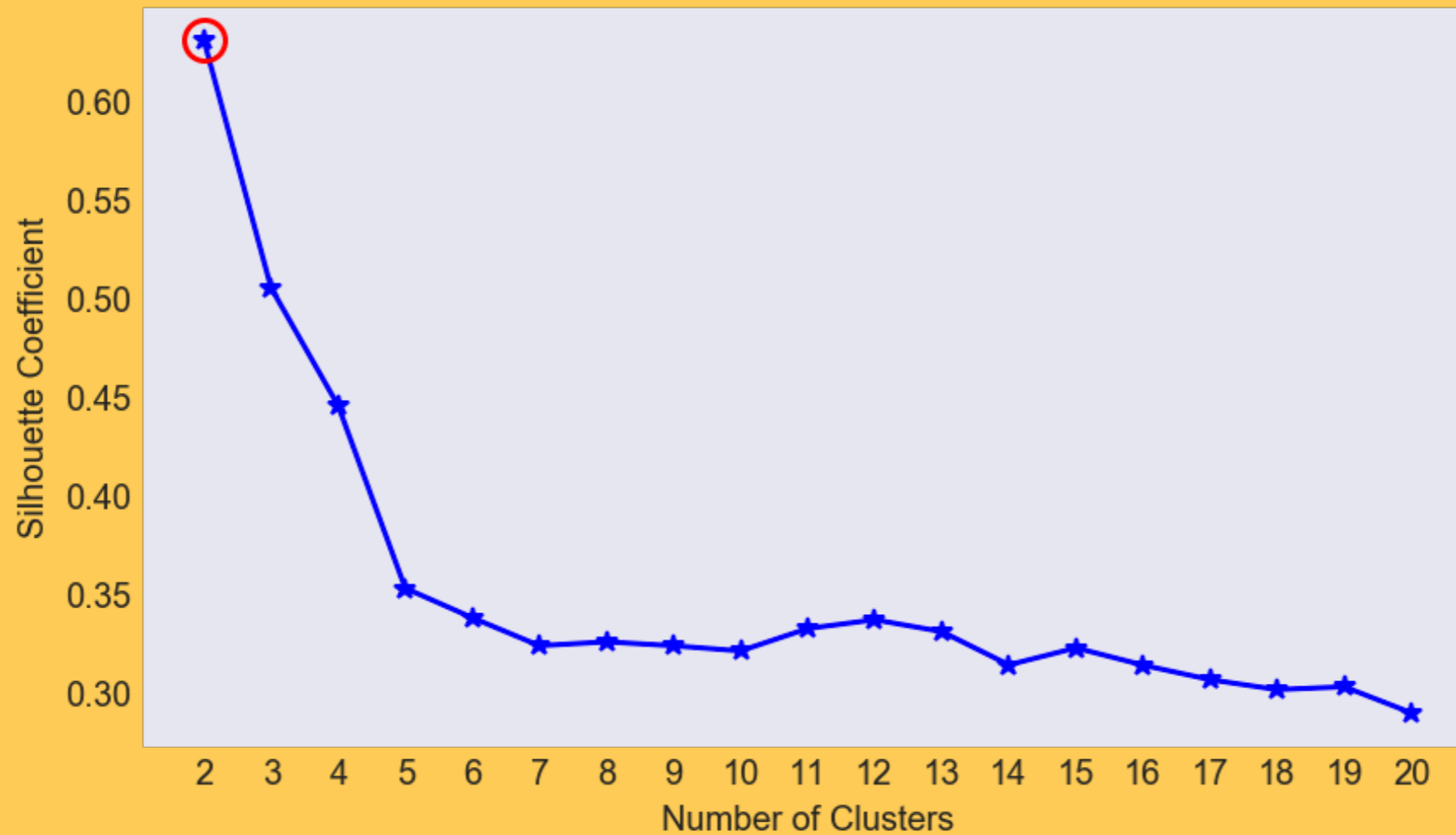
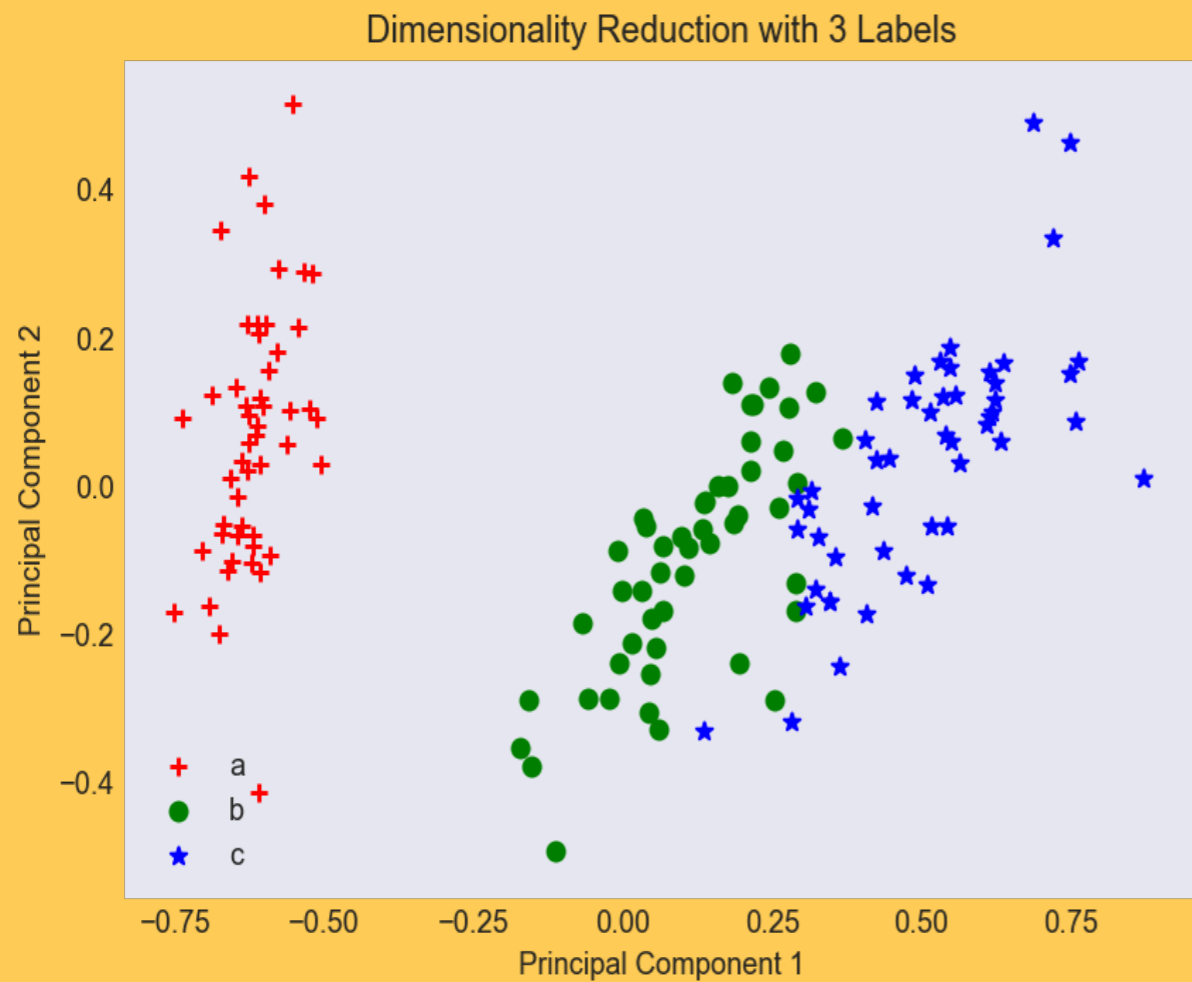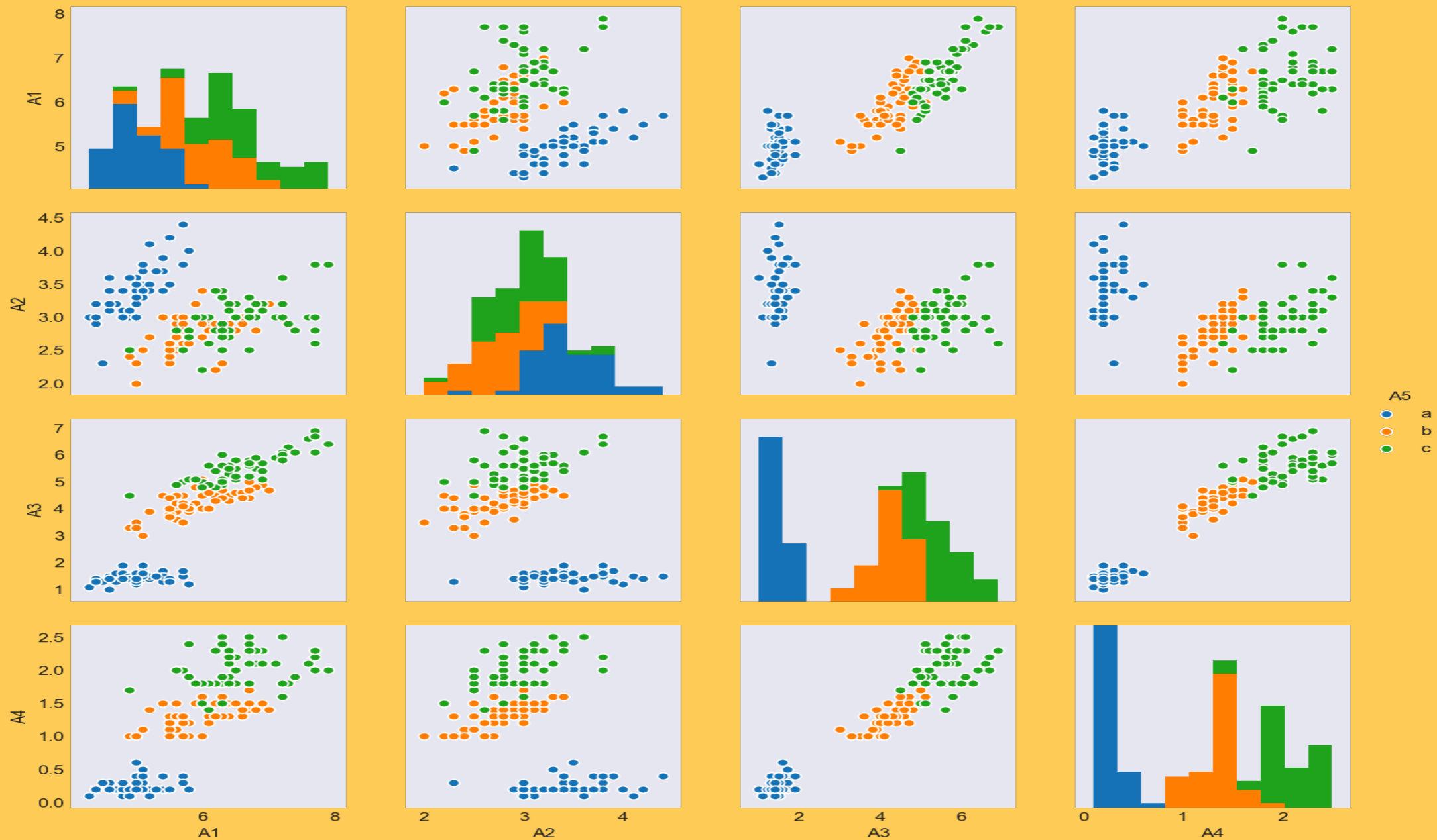# B AND C ARE GROUPED TOGETHER

3 CENTROIDS IS NOT A LOT BETTER...

DBSCAN BETTER BUT HAS OUTLIERS

Dimensionality Reduction with 3 Labels

Density Scan reveals 3 Clusters and outliers
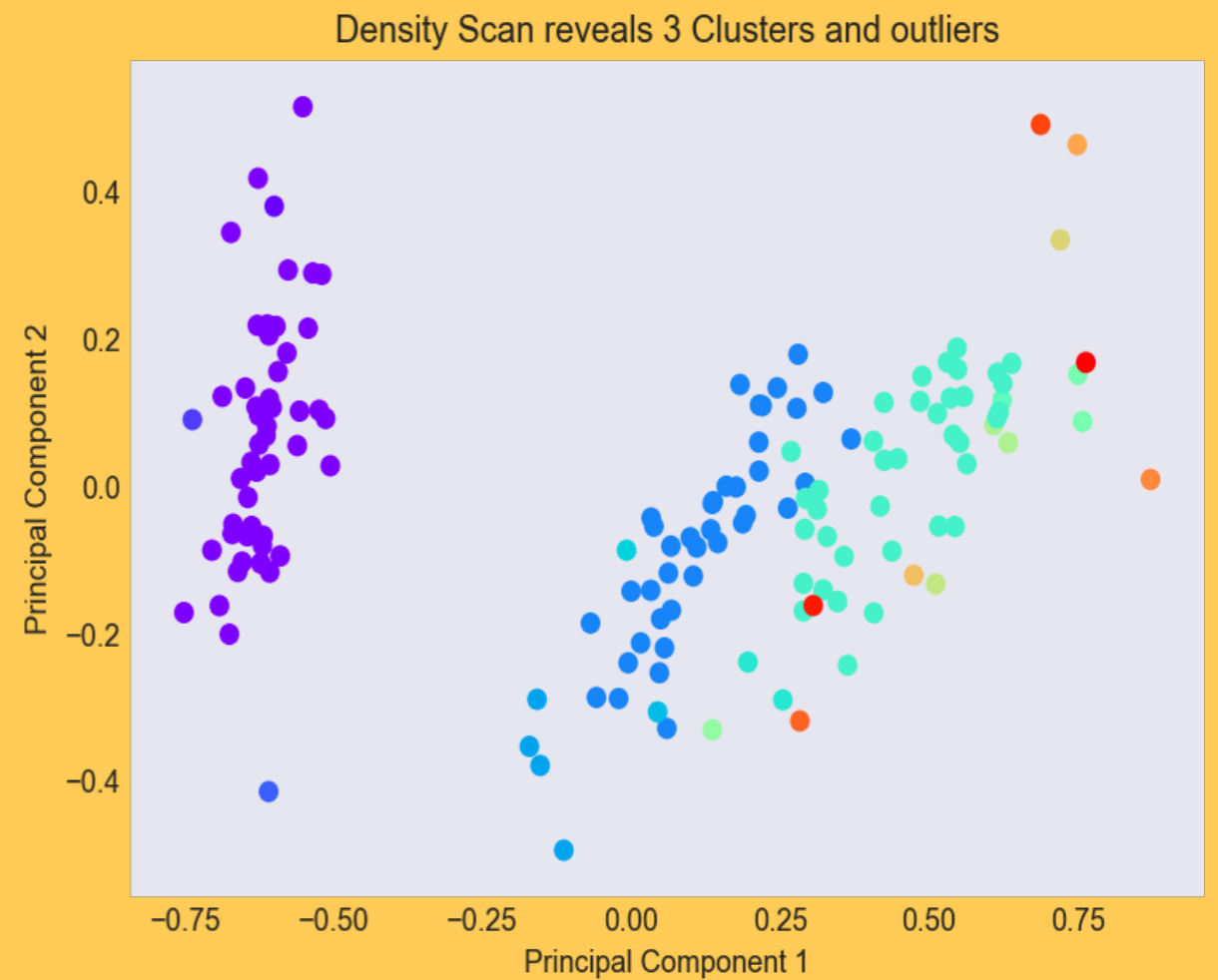
WHY K-MEANS DID POORLY

SECTION 2

LOGISTIC CLASSIFIER

▸ The more young children you are looking after and the more hungry you are, but the less alternative eating options in your area means a higher probability of you eating dinner at McDonalds tonight.
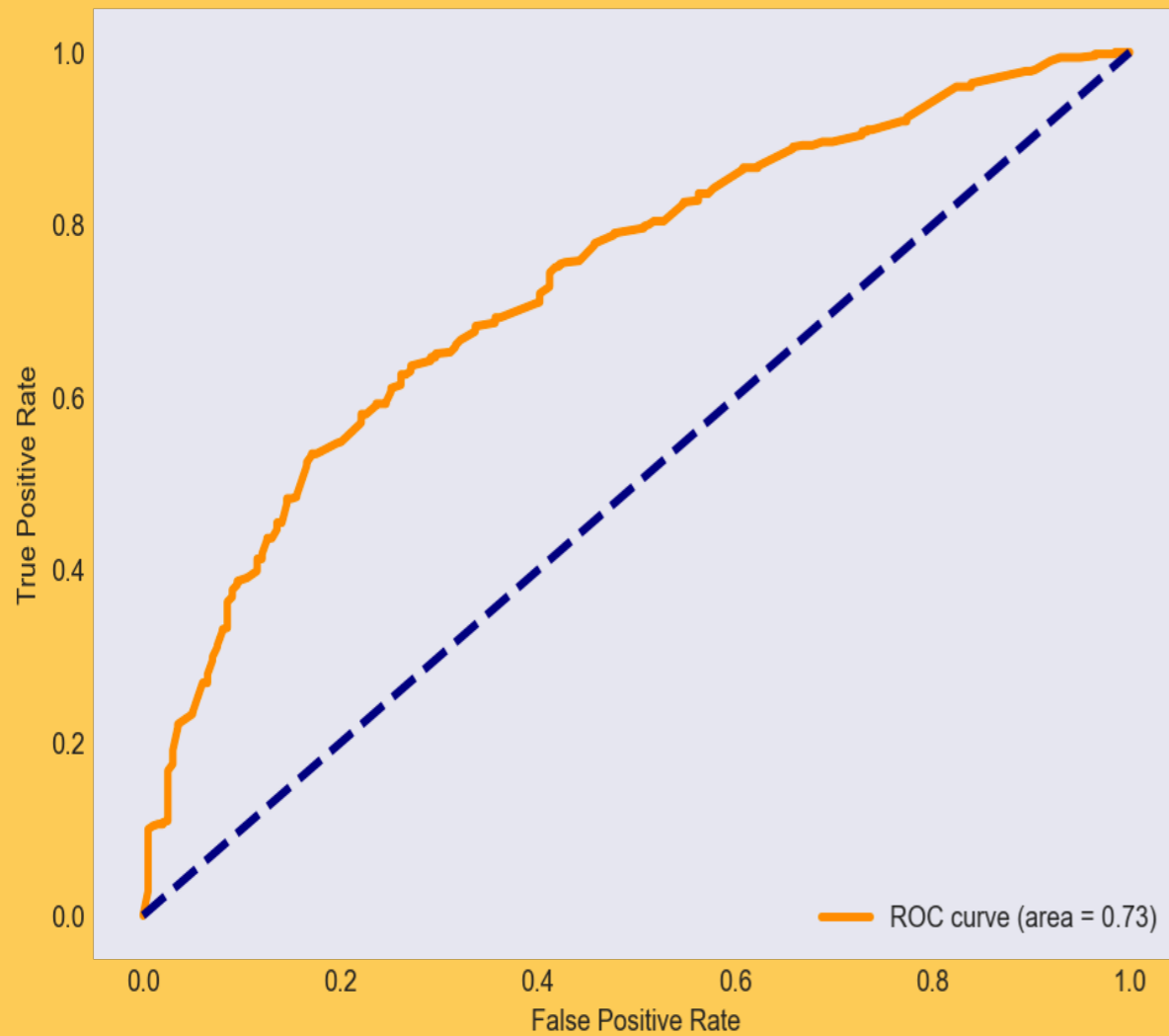
▸ More comprehensible model

BEST FEATURES

▸ Odds of getting into Class 1 over Class 0

▸ Holding other features constant

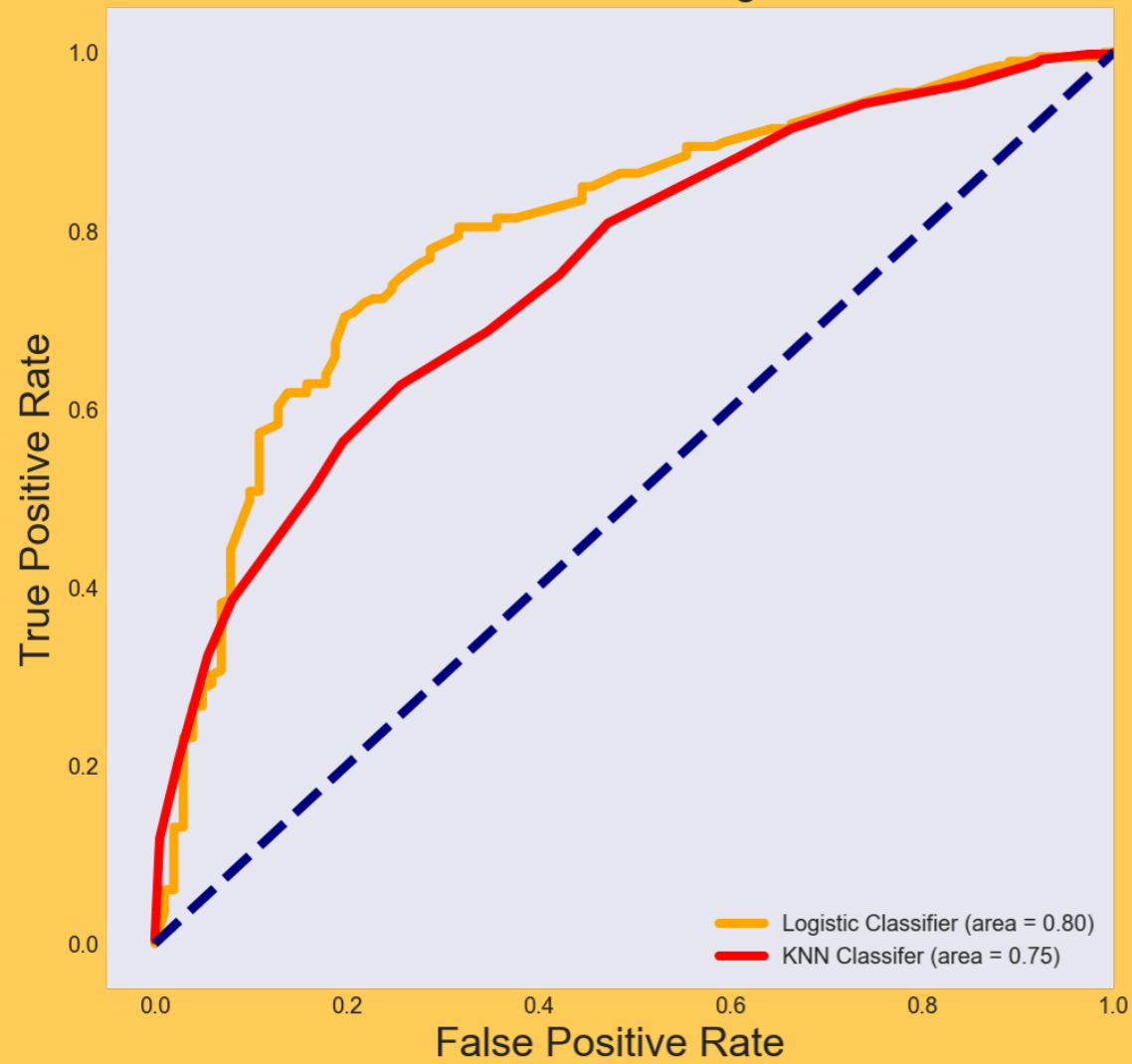| Feature	Coefficient	Odds
A6	1.124984	3.08
A7	0.453731	1.57
A9	0.364034	1.44
A10	0.113566	1.12

70-30 TRAIN-TEST SPLIT

KNN CLASSIFIER

▸ Residents of Glebe like residents of Newtown, Marrickville, Erskineville and Surry Hills are more likely to take public transport to the Central Business District

▸ Group into a class based on neighbors

▸ Also intuitive

70-30 TRAIN-TEST SPLIT

FOUR FACTOR MODEL