# Uncovering NFL Player Archetypes Using Clustering

CSCI-B 365 - Data Mining Project

Emily Ahern, Sydney Gargiulo, Jack Schwartz, Noah Bussell

# Introduction

- Applies clustering techniques to NFL player data to uncover distinct football archetypes based on performance and physical attributes.
- By leveraging K-means clustering, Principal Component Analysis (PCA), and cluster evaluation methods such as the Elbow and Silhouette scores, the objective is to identify meaningful patterns within the data.
- These clusters provide insights into the different skill sets and play styles that define various player types, offering valuable perspectives for scouting, performance evaluation, and strategy development.

# Dataset Description

- Source: Kaggle NFL Combine Dataset (2000–2022)
- Features Used:
  - 40yd dash: (4.26 - 5.29) sec
  - Vertical: (25 – 46.5) in
  - Bench: (3 – 34) Reps of 225 lbs
  - Broad Jump: (98 – 140) in
  - 3Cone: (6.28 - 7.96) sec
  - Shuttle: (3.75 - 4.96) sec
  - Height: (65 – 80) in
  - Weight: (168 – 336) lbs
- Data Size: 1489 players across 7 major positions (WR, RB, CB, TE, LB, S, QB)

# Methodology

- Data Mining Technique: Clustering
  - Applied to NFL dataset to group players based on their similar attributes
- Data Preprocessing
  - Data is cleaned; dropping NAs and unnecessary attributes
  - Features are scaled to ensure balance in analysis
- K-means Clustering


- Silhouette Score & Elbow Method
  - Gives the optimal number of clusters according to the data, balancing intra-cluster compactness and inter-cluster separation
- Cluster Interpretation and Analysis
  - Used PCA plots, radar maps, and heat maps to visualize the shared attributes among clustered NFL players

# Brief Timeline

## Week of April 21st

- Data Preprocessing
- Initial Planning

## Week of April 28th

- Perform K-means clustering
- Complete code
- Create layout of our final analysis

## May 3rd – May 6th

- Complete final presentation
- Complete project report
- May 5th: Presentation

# Preprocessing Steps

- Combined the datasets from the years 2000-2022
- Converted height to inches
- Filtered relevant positions
- Removed rows with missing key combine stats
  - (3895, 13), dropped 304 rows
- Scaled numerical values using StandardScaler

Combining data

```
# Load and combine all years
combine_dfs = []
for file in combine_files:
    df = pd.read_csv(file)
    df["Year"] = int(os.path.basename(file).split("_")[0])  #
    combine_dfs.append(df)

combine_df = pd.concat(combine_dfs, ignore_index=True)
```

```
# === Step 1: Load and Prepare Data ===
df = pd.read_csv("cleaned_combine.csv")
combine_metrics = ["40yd", "Vertical", "Bench", "Broad Jump", "3Cone", "Shu
df_clustering = df[combine_metrics].dropna()
```

```
# === Step 2: Convert Height to Inches ===
def convert_height(ht):
    try:
        feet, inches = map(int, ht.split('-'))
        return feet * 12 + inches
    except:
        return np.nan

combine_df["Height_in"] = combine_df["Ht"].apply(convert_height)
```

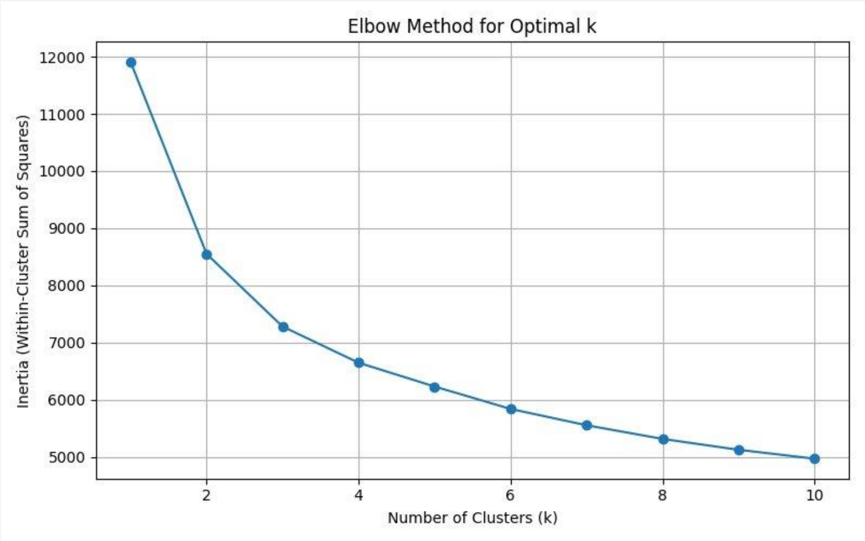
```
# === Step 3: Filter to Relevant Positions ===
relevant_positions = ['RB', 'WR', 'CB', 'LB', 'TE', 'QB', 'S']
combine_df = combine_df[combine_df["Pos"].isin(relevant_positions)]
```

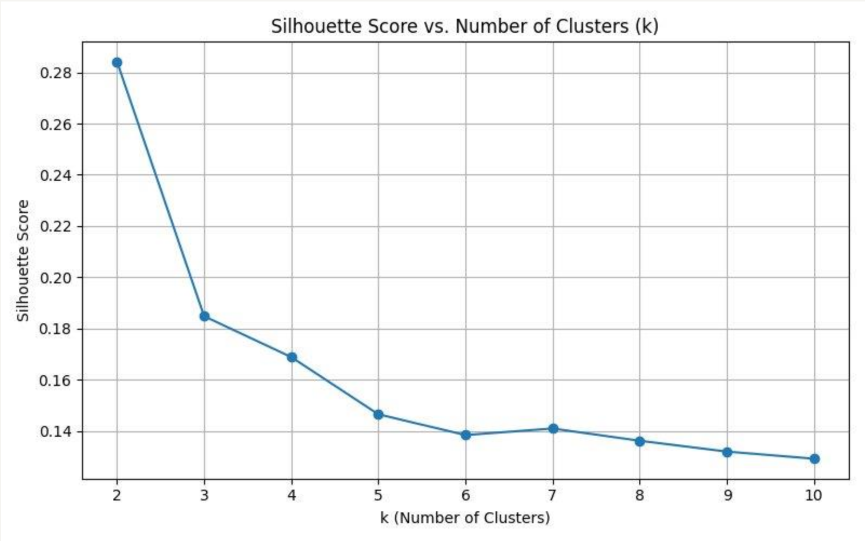
```
# === Step 2: Normalize the Data ===
scaler = StandardScaler()
X_scaled = scaler.fit_transform(df_clustering)
```

# Elbow & Silhouette Evaluation

- Elbow Curve: Optimal bend near k = 2
- Silhouette Scores: Highest at k = 2
- Interpretation: Tight intra-cluster and good inter-cluster separation at k = 2

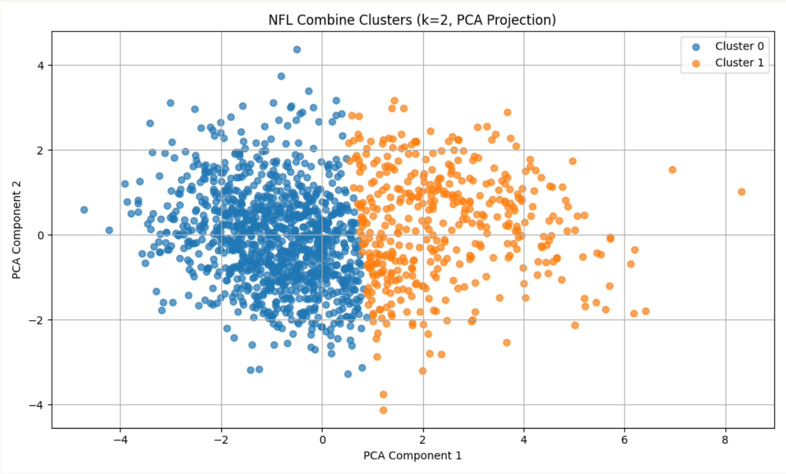
K=2, Silhouette Score = 0.2842
K=3, Silhouette Score = 0.1848
K=4, Silhouette Score = 0.1688
K=5, Silhouette Score = 0.1465

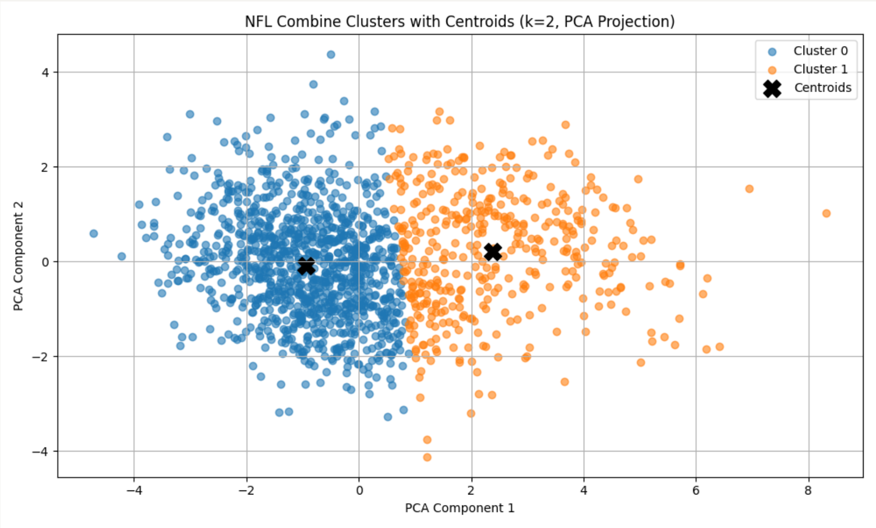


# PCA Cluster Visualization (k=2)

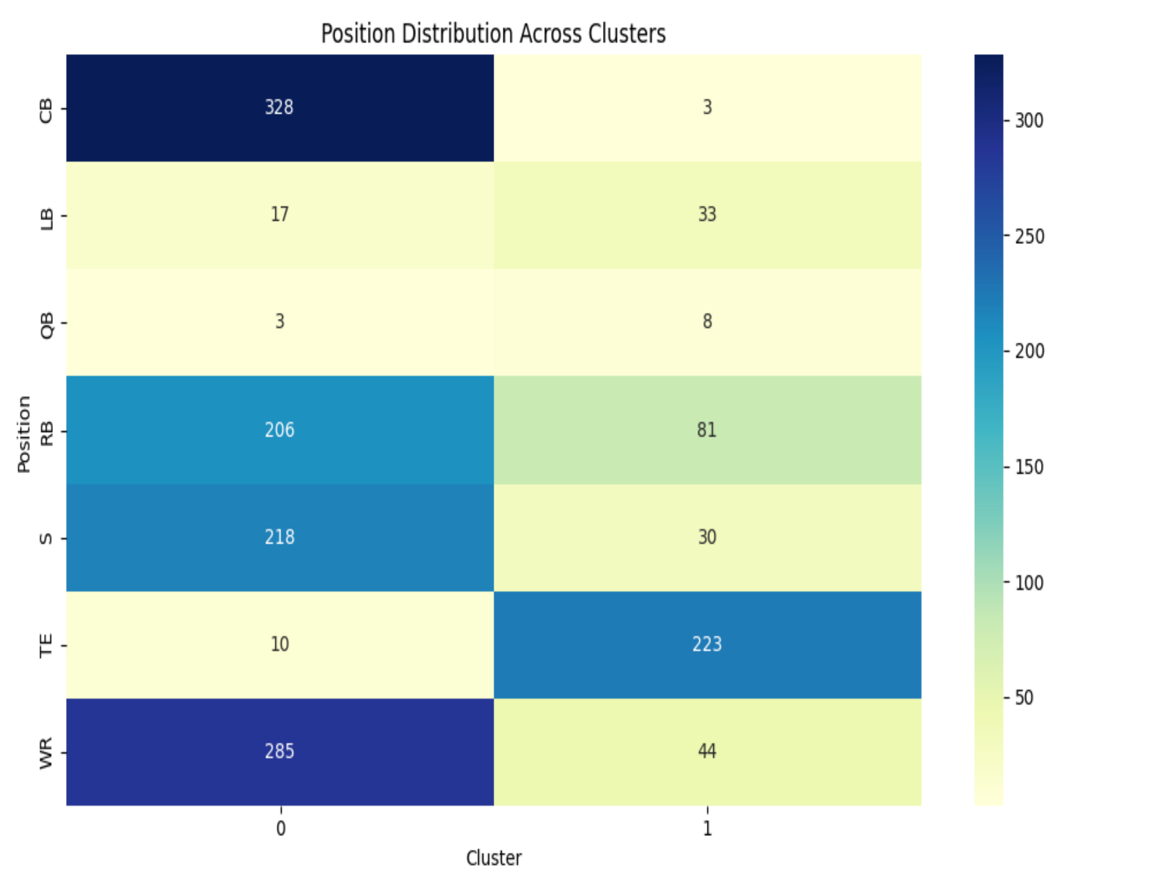
- 2D scatter plot of clusters
- Black 'X' markers show centroids
- Clear separation visible via PCA



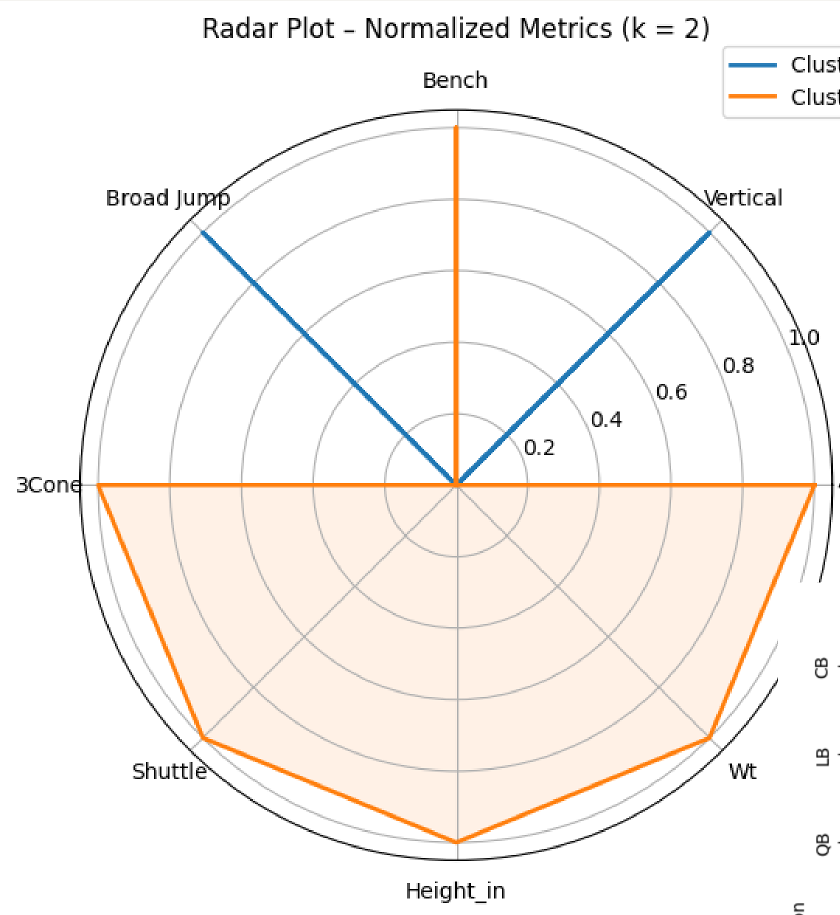



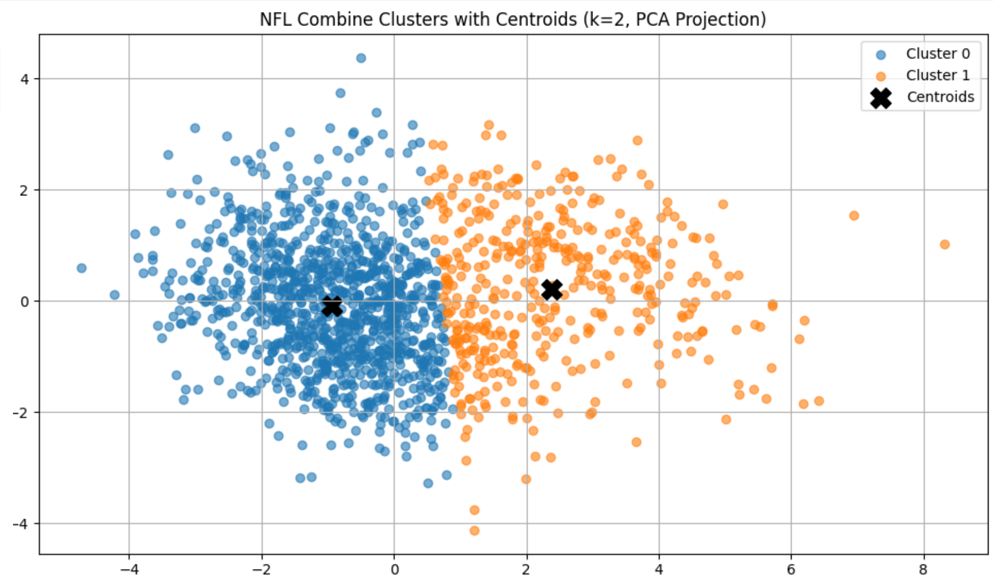
# Position Distribution by Cluster (k=2)

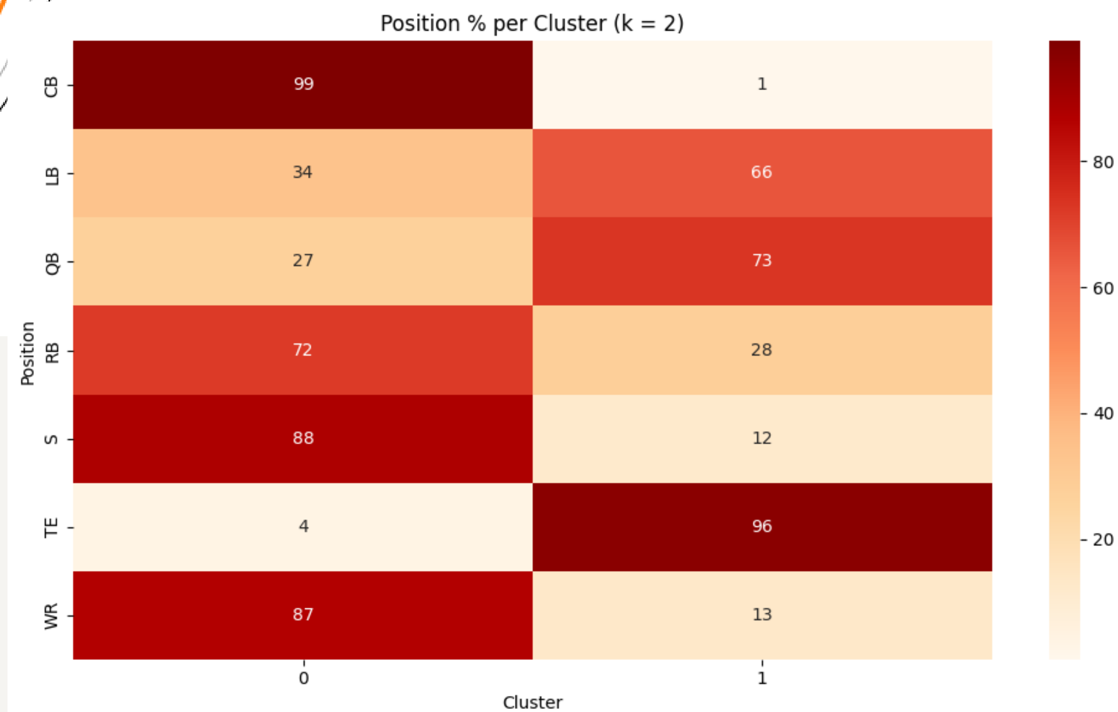
- Heatmap showing which positions dominate each cluster
- Cluster 0: WR, CB, S, RB (faster, more agile)
- Cluster 1: TE, LB, QB (bigger, stronger)



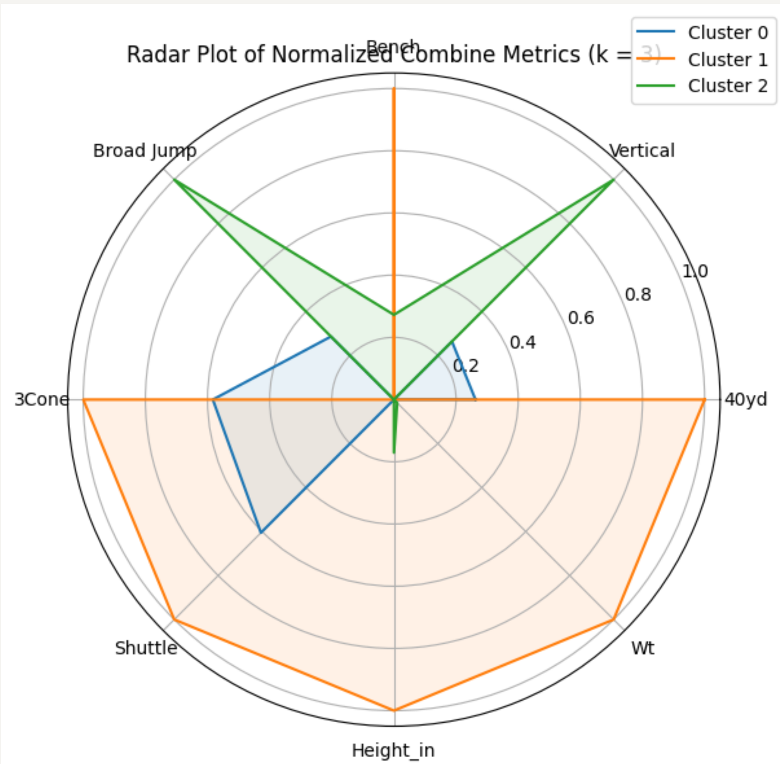
# K-Means Clustering for k = 2

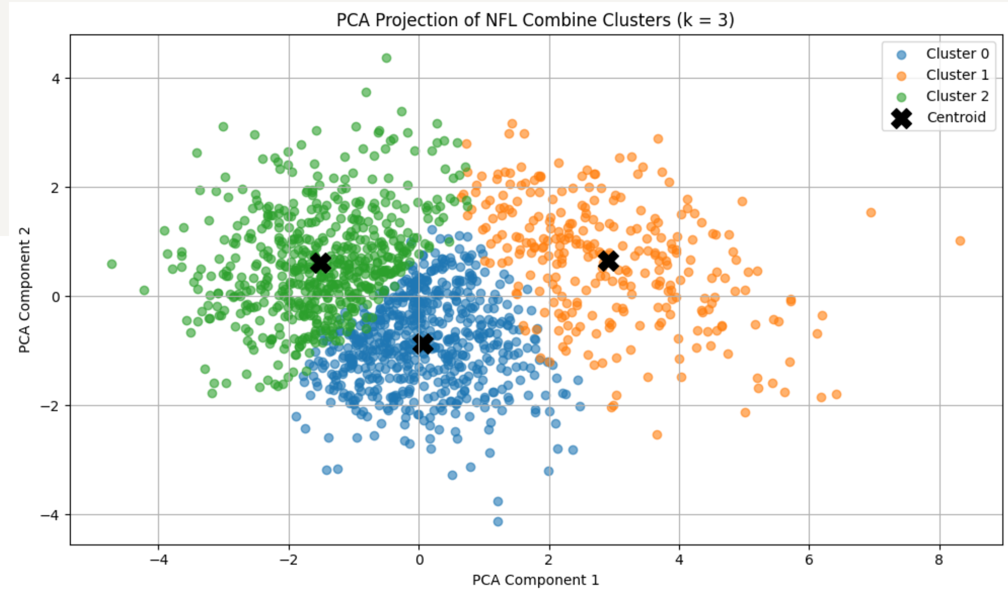


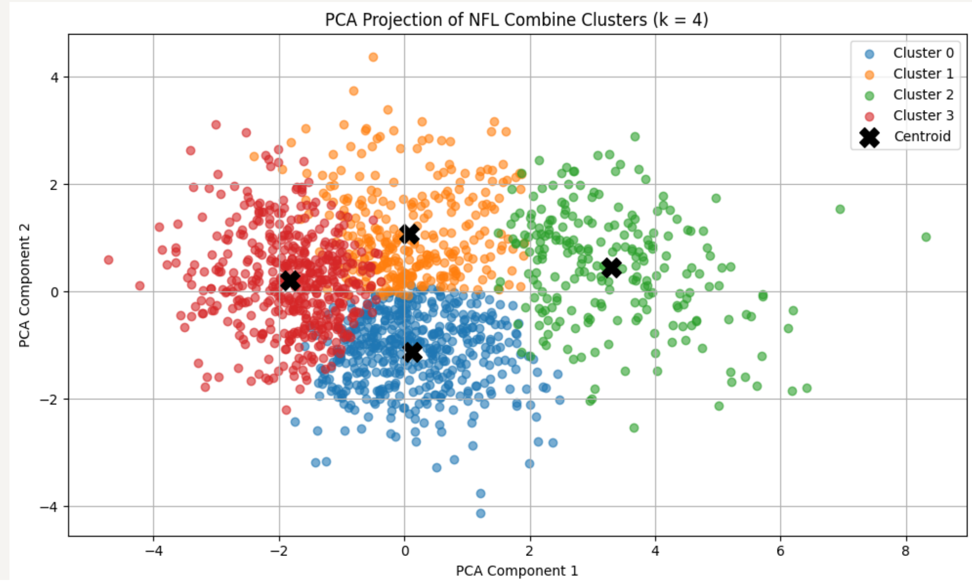



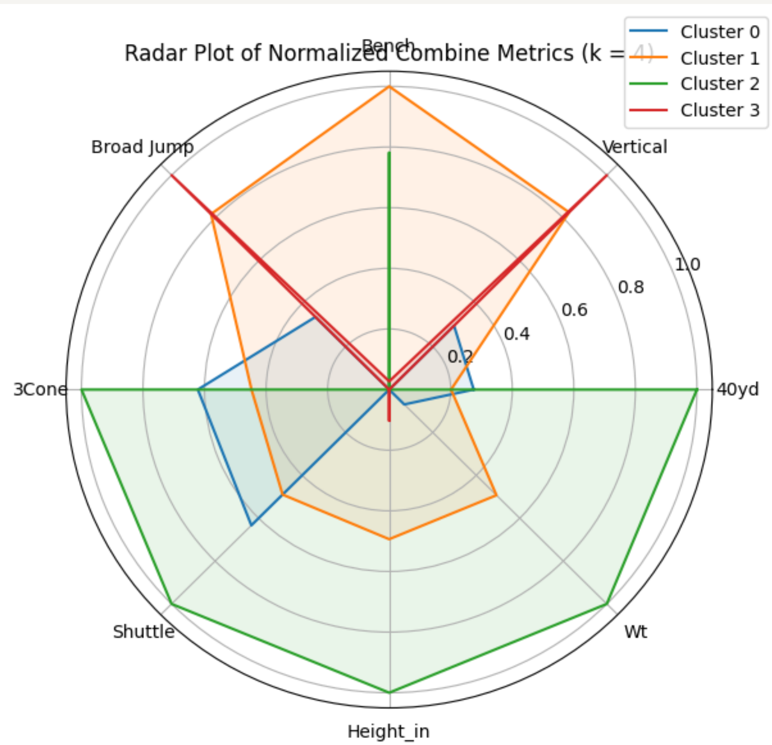
# Same Evaluation for k = 3

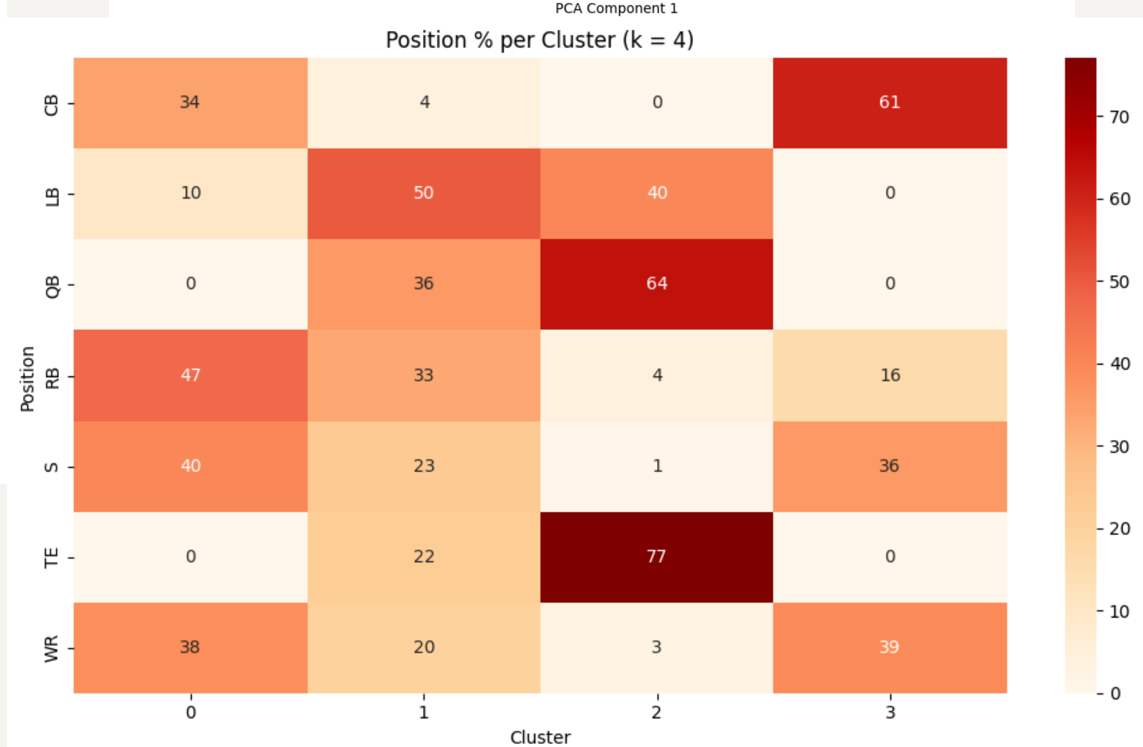


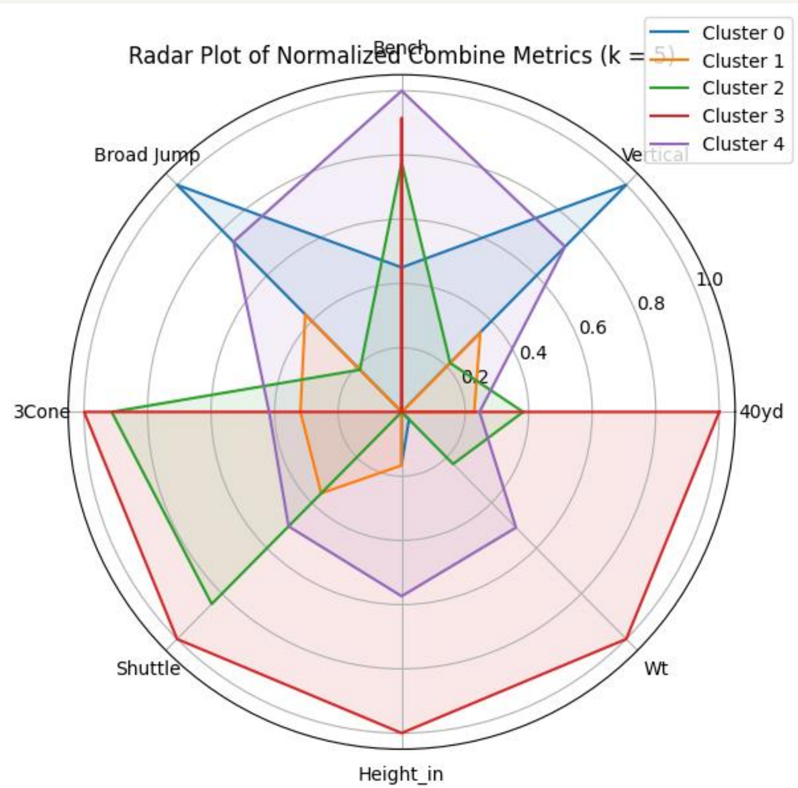

# Same Evaluation for k = 4

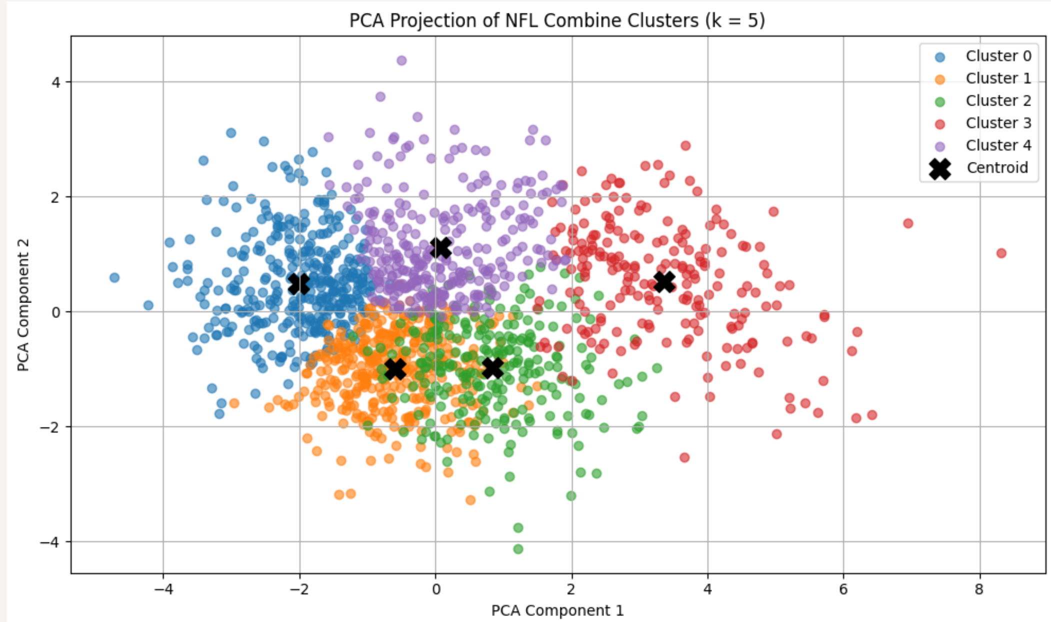


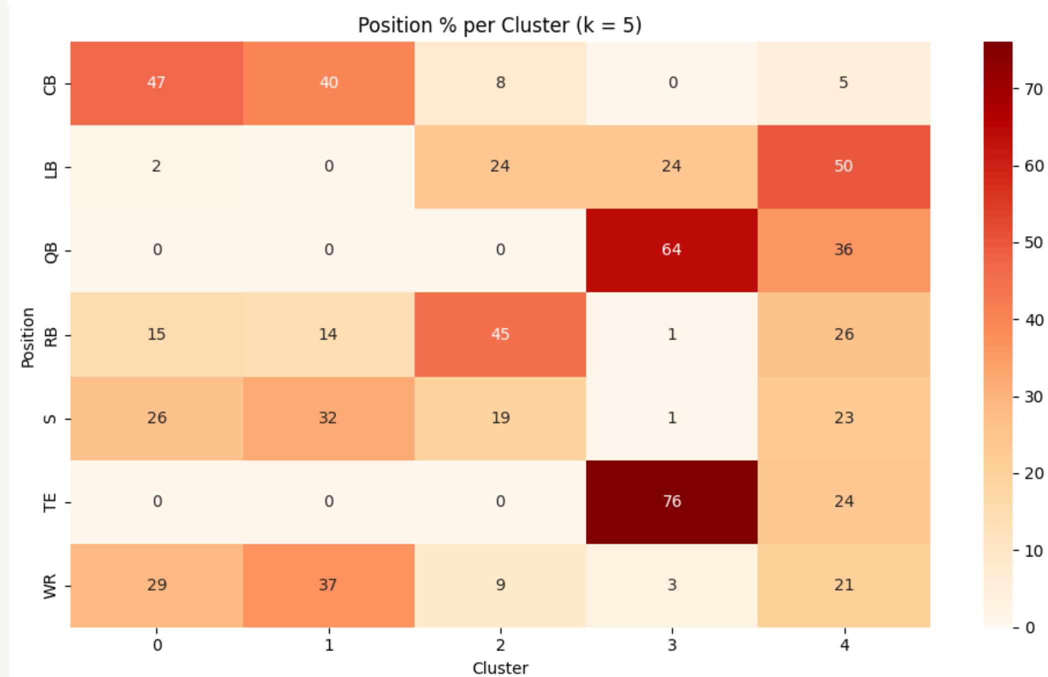



# Same Evaluation for k = 5



# Key Findings

- Clustering was most efficient with k =2
- Clustering successfully revealed athletic archetypes
- Position distribution aligned with cluster traits
- Visual tools like PCA were key to interpreting results

# Thank you, any further questions?