Clustering – Sporting Cristal

Futbol Analytics Club – Ohio University

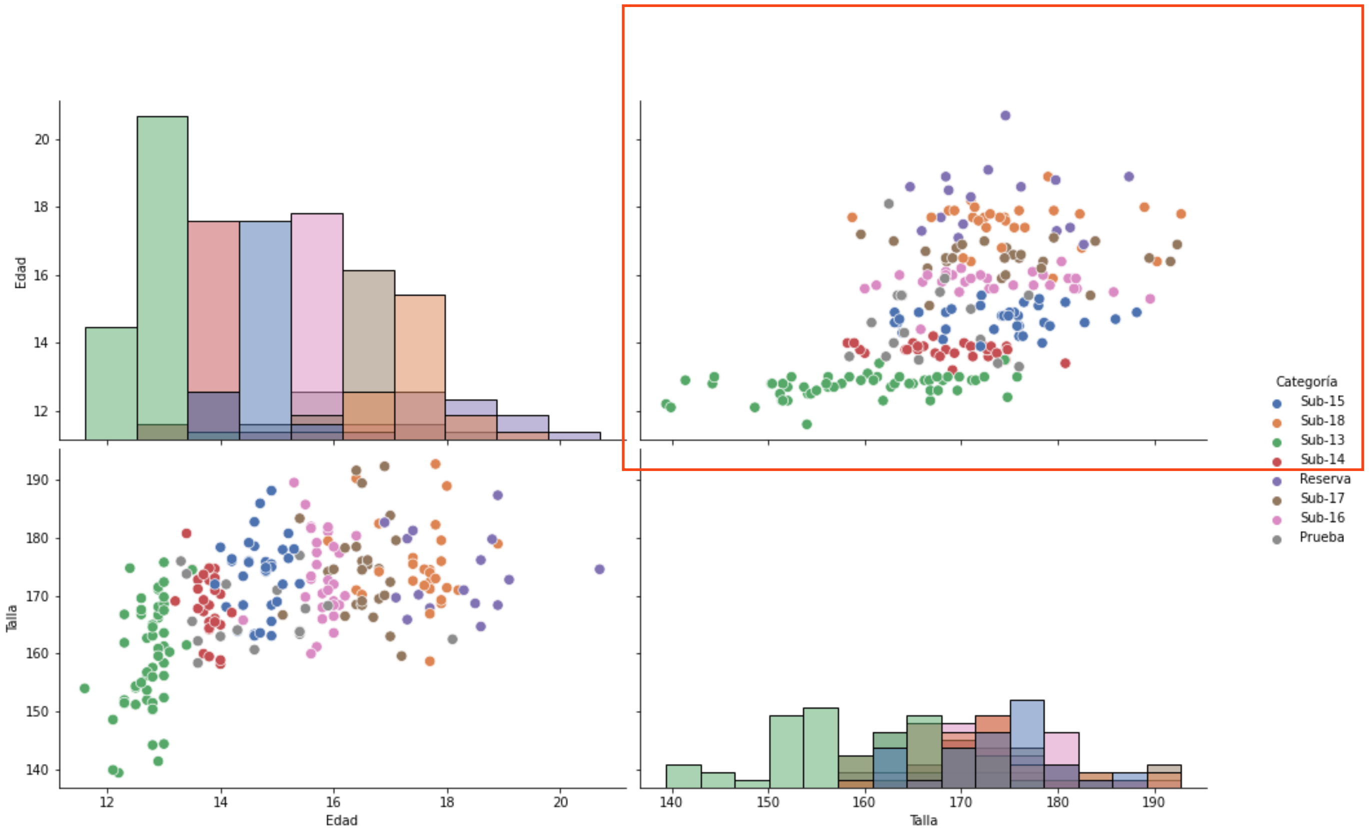
Task 01

The club is currently utilizing anthropometric data from various academy and reserve team players to evaluate their performance in terms of different body composition variables, such as 'Masa Adiposa' (Body Fat Mass), 'Masa Muscular' (Muscle Mass), and 'Suma de Pliegues' (Sum of Skinfolds).

However, the current methodology involves classifying and comparing players within the same age category, leading to potential biases and inaccuracies. For instance, a 15-year-old player may be grouped with 18-year-olds, despite significant differences in physical development. Height also plays a crucial role, as taller individuals naturally tend to have higher muscle mass, for example.

Therefore, it's essential to devise a more nuanced classification system that accounts for both age and height. By doing so, we aim to ensure that team analyses and assessments accurately reflect the players' morphological and biological realities

Visual representations, such as scatterplots, are employed to depict the data, with colors indicating player categories. However, considerable overlap suggests the need for refinement in the classification approach.

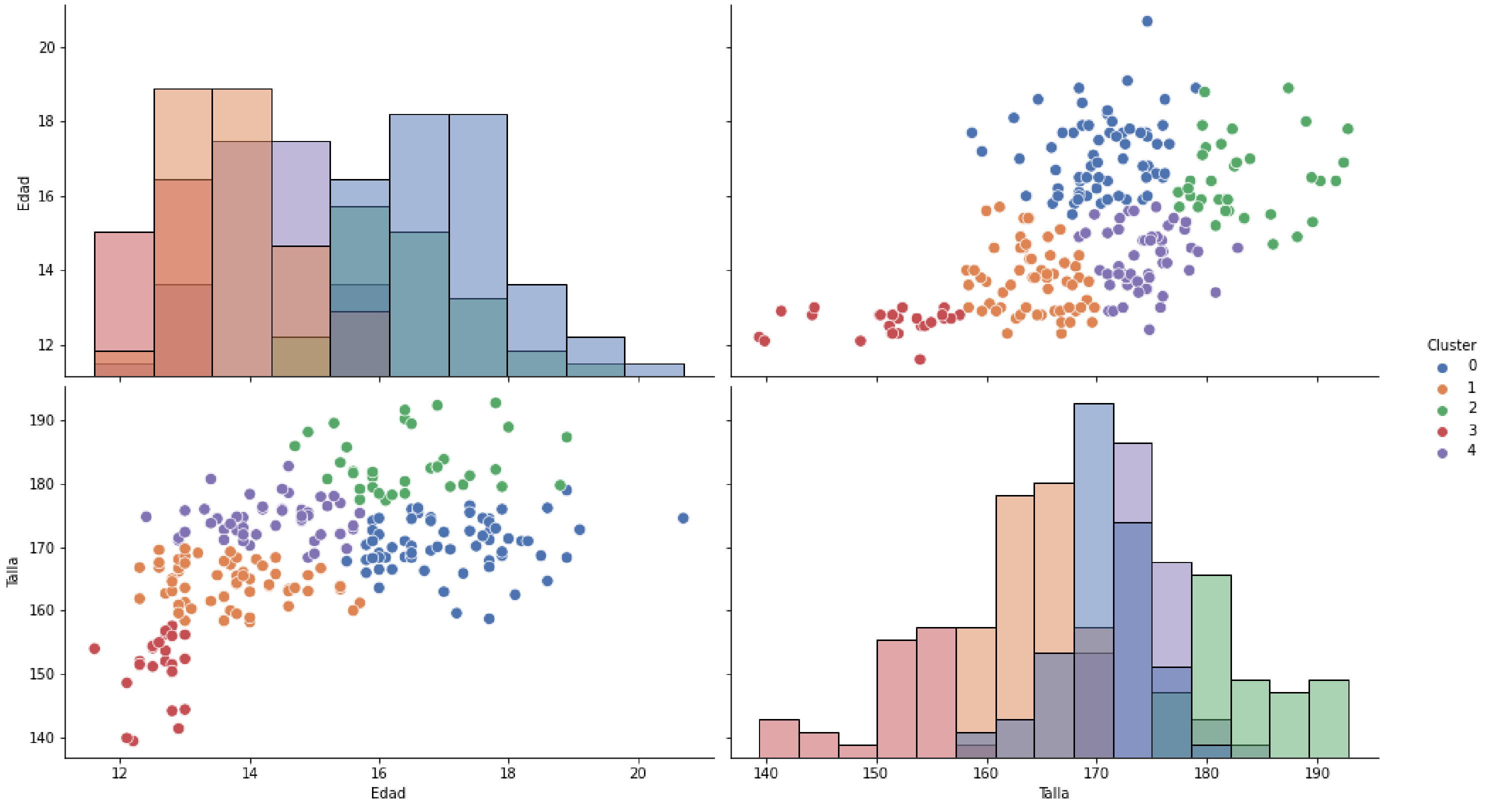


Kmeans (n_clusters=**5**) | Kmeans_model.intertia_ = **92.51**

5 Clusters

We employed an unsupervised machine learning technique known as K-means clustering to automatically group players according to their key characteristics: 'Height' and 'Age'. The goal was to uncover underlying patterns and structures within the unlabeled data and organize players based on their similarities.

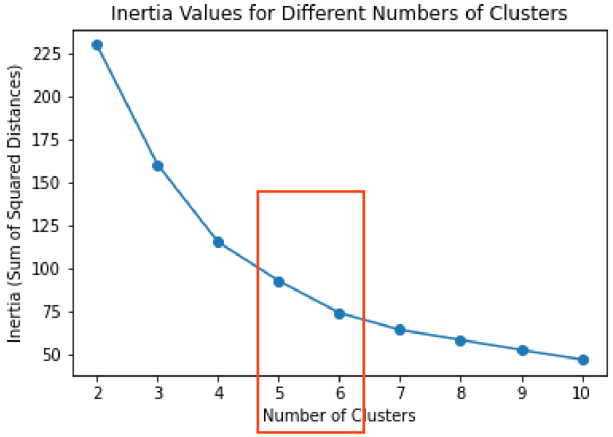
To determine the optimal number of clusters, we utilized the Elbow method, as detailed in the Jupyter Notebook provided for reference. After analysis, we decided to focus on exploring 5 and 6 clusters to gain insights into the data's organization and distribution.

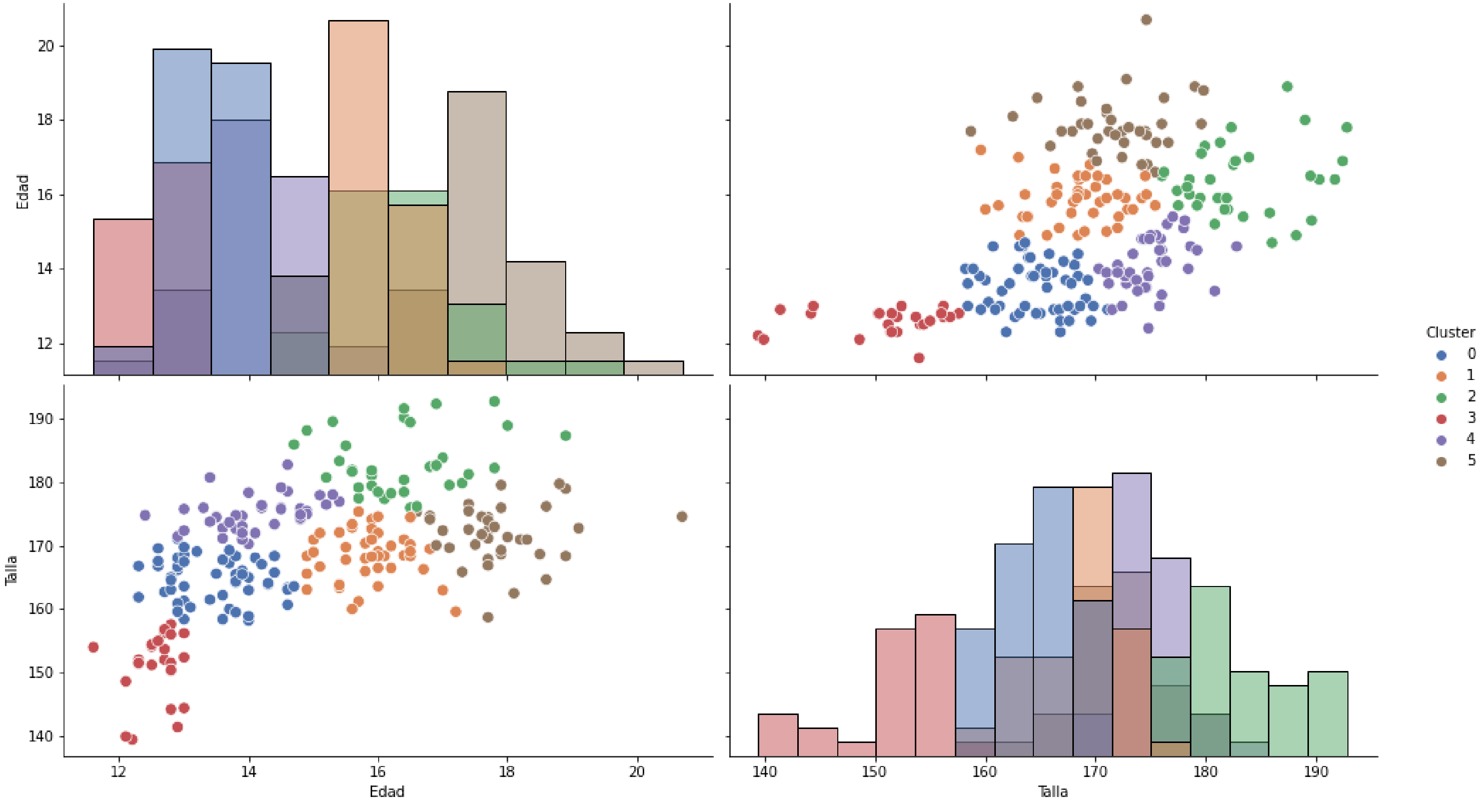


Kmeans (n_clusters=**6**) | Kmeans_model.intertia_ = **73.86**

6 Clusters

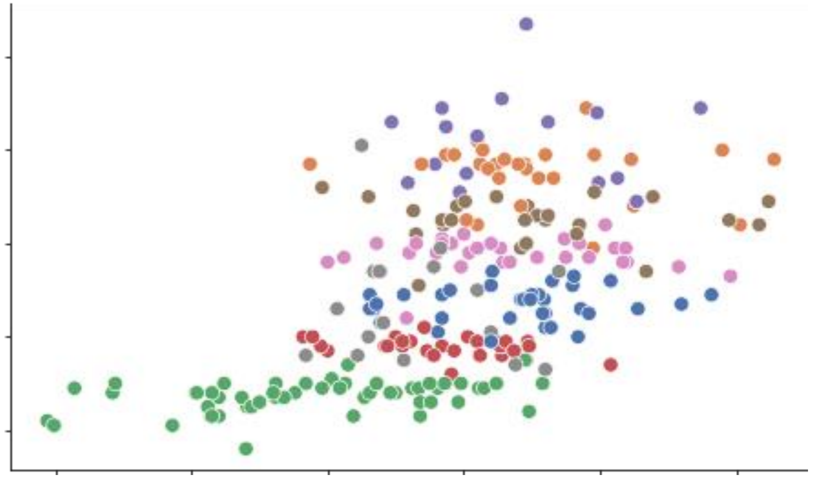
Similarly, we opted to use 6 clusters for our analysis. In both the 5-cluster and 6-cluster scenarios, we observed improved organization in the data within the scatterplot. However, the decision on which clustering configuration to adopt moving forward will rely on additional analysis and domain expertise. Further details on this matter will be provided in the following slides.



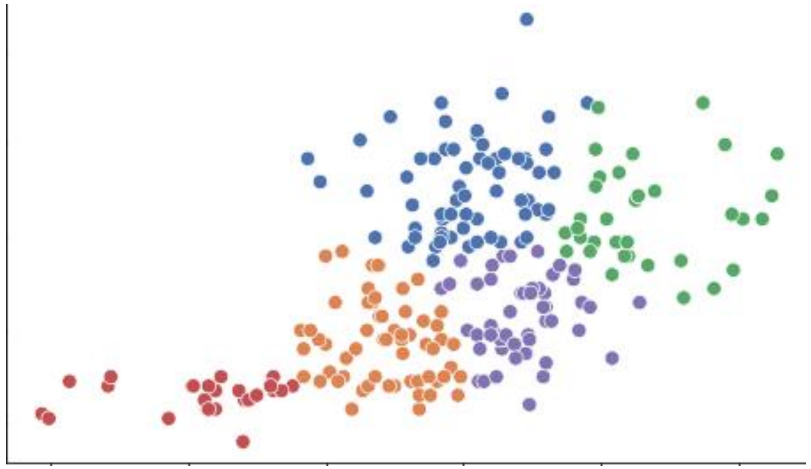


Visual Comparison

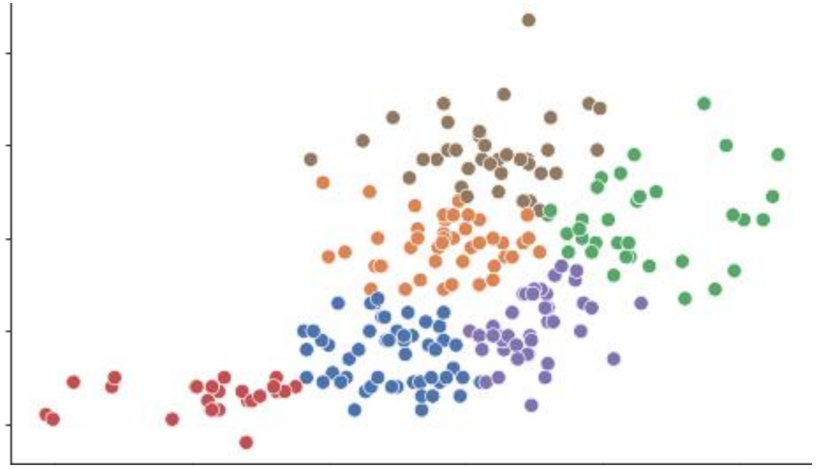
By Category



5 Clusters



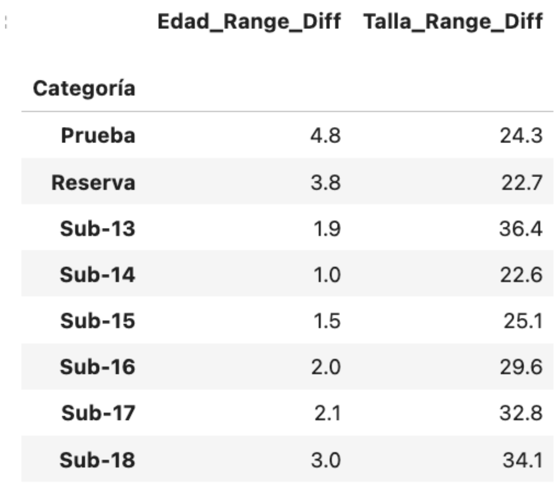
6 Clusters



Output Analysis

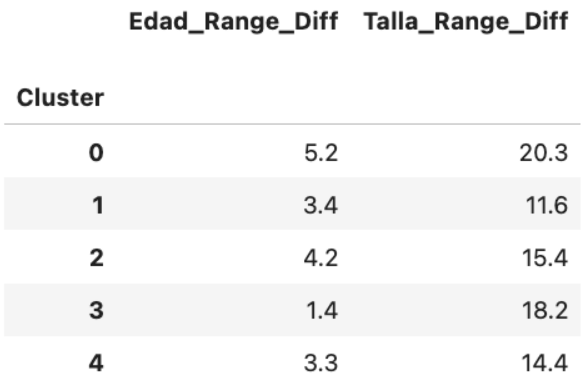
The K-means model has demonstrated its effectiveness in clustering, particularly when evaluating the range differences per cluster for both Age and Height, in comparison to the existing practice ranges. Significantly improved results were observed, particularly in terms of Height, which is a pivotal variable in assessing body composition structure.

Current



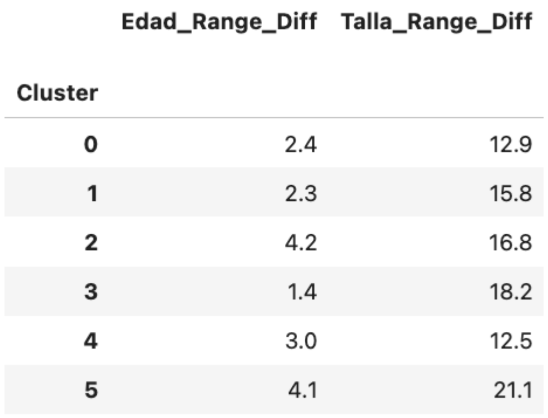
	Edad_Range_Diff	Talla_Range_Diff
Categoría		
Prueba	4.8	24.3
Reserva	3.8	22.7
Sub-13	1.9	36.4
Sub-14	1.0	22.6
Sub-15	1.5	25.1
Sub-16	2.0	29.6
Sub-17	2.1	32.8
Sub-18	3.0	34.1

5 Clusters



	Edad_Range_Diff	Talla_Range_Diff
Cluster		
0	5.2	20.3
1	3.4	11.6
2	4.2	15.4
3	1.4	18.2
4	3.3	14.4

6 Clusters



	Edad_Range_Diff	Talla_Range_Diff
Cluster		
0	2.4	12.9
1	2.3	15.8
2	4.2	16.8
3	1.4	18.2
4	3.0	12.5
5	4.1	21.1

Further Analysis

Now, let's compare the three outputs based on a similar 'Age' index.

- Specifically, we'll focus on the **Sub-17** category, where the mean age is **16.50**.
- In the 5-cluster scenario, we'll examine **Cluster 2** with an age mean of **16.47**.
- In the 6-cluster scenario, we'll also analyze **Cluster 2** with an age mean of **16.37**.

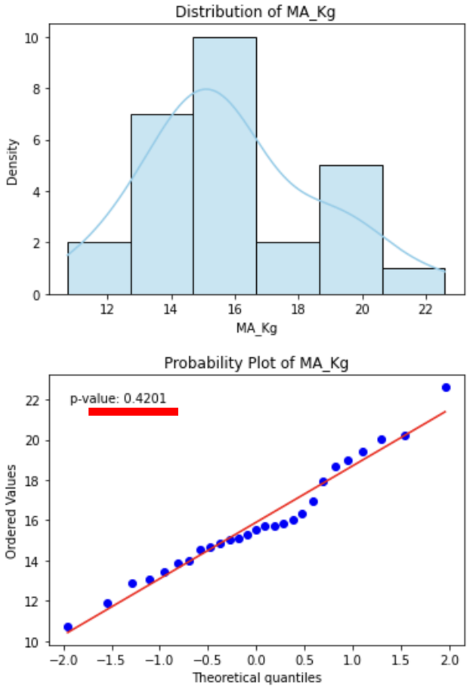
Grouping	Average Age	Average Height
Sub-17	16.5 (std: 0.48)	174.5 (std: **8.33**)
Cluster 2 (5C's)	16.47 (std: 1.03)	183.3 (std: 4.52)
Cluster 2 (6C's)	16.37 (std: 0.91)	183.1 (std: 4.75)

A smaller standard deviation indicates less variability in the distribution of the data. When using Height as the main variable for clustering, a significant reduction in standard deviation—from 8.33 to 4.75, for instance—indicates a more compact and distinct clustering of data points based on their height characteristics.

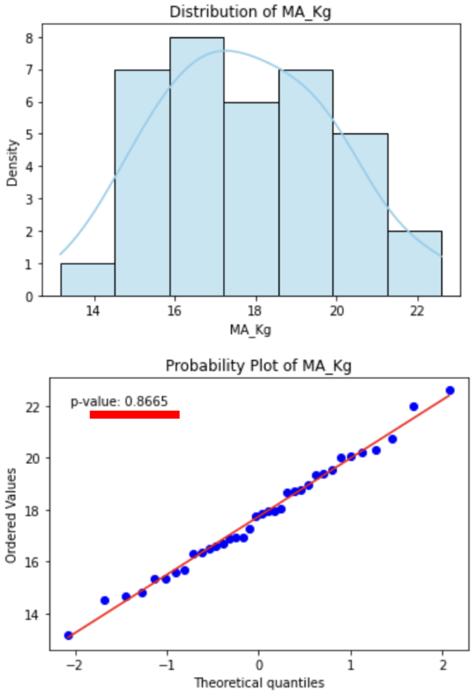
Target Variable – Masa Adiposa (Kg)

We now visualize the frequency and distributions of the target variable values. The p-values and probability plot suggest stronger evidence that the clusters provide a normally distributed pattern. Additionally, there is compelling evidence indicating that the distribution of muscular mass under the Sub-17 category does not follow a normal distribution, as indicated by its p-value (<0.05).

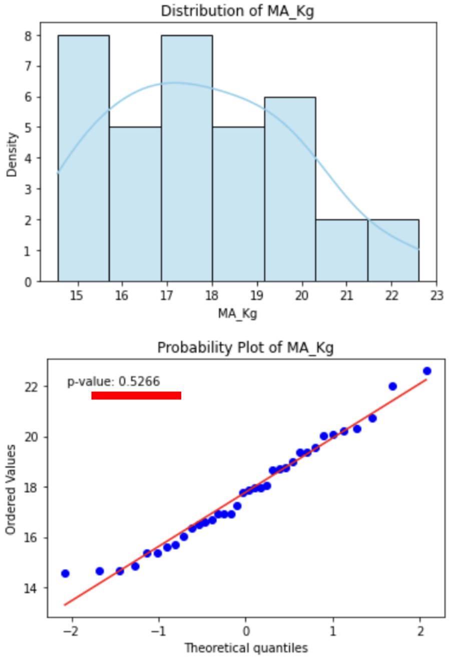
Sub-17



Cl. 2 (5C's)



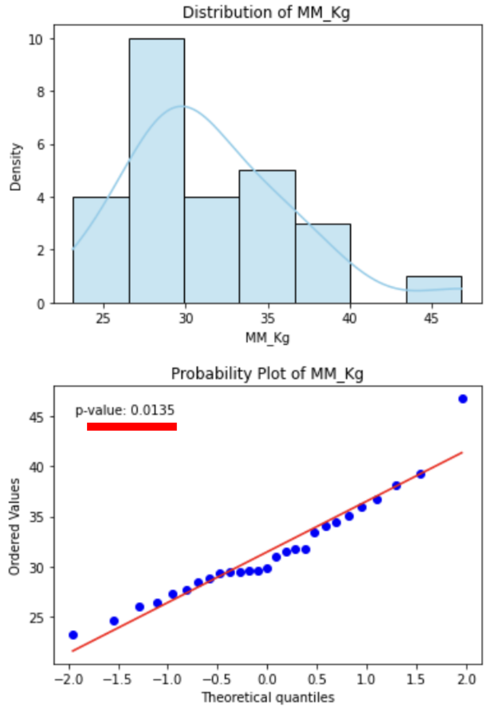
Cl. 2 (6C's)



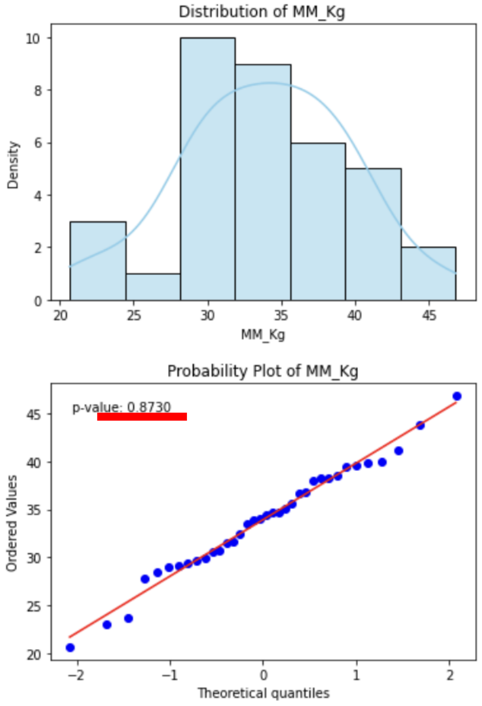
Target Variable – Masa Muscular (Kg)

We now visualize the frequency and distributions of the target variable values. The p-values and probability plot suggest stronger evidence that the clusters provide a normally distributed pattern. Additionally, there is compelling evidence indicating that the distribution of muscular mass under the Sub-17 category does not follow a normal distribution, as indicated by its p-value (<0.05).

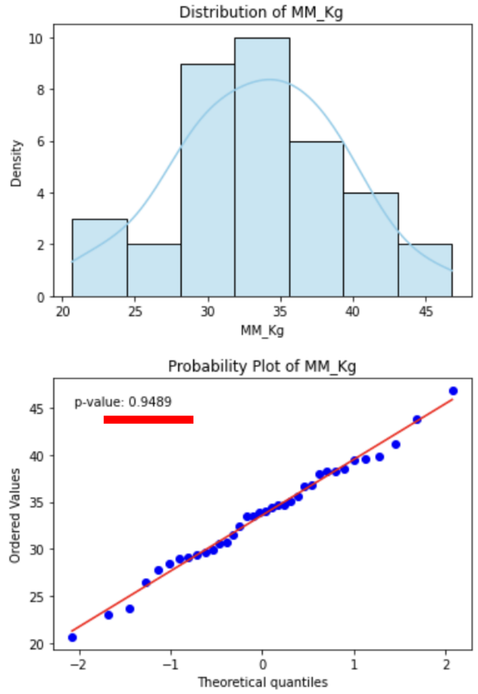
Sub-17



Cl. 2 (5C's)



Cl. 2 (6C's)



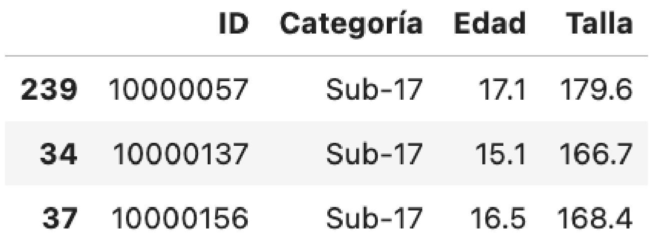
Comparing Random Players Within the Grouping

The objective here is to explore the data by selecting three random players within each grouping: Sub15, Cluster 2 (from 5 clusters), and Cluster 2 (from 6 clusters).

The main question to consider is: 'Does it make sense to compare these 3 players?'

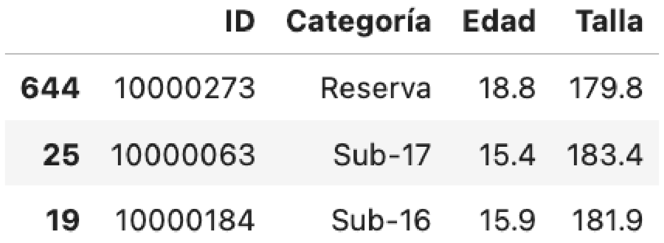
It appears that both Cluster 2 groups incorporate both variables, rather than solely categorizing players based on the category they play

Sub-17



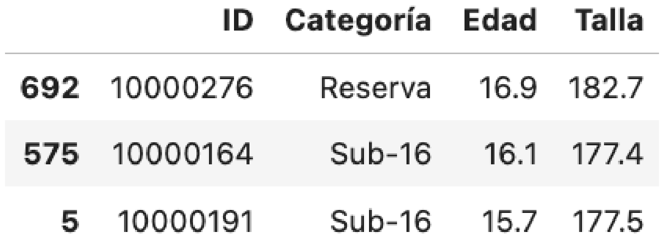
	ID	Categoría	Edad	Talla
239	10000057	Sub-17	17.1	179.6
34	10000137	Sub-17	15.1	166.7
37	10000156	Sub-17	16.5	168.4

Cl. 2 (5C's)



	ID	Categoría	Edad	Talla
644	10000273	Reserva	18.8	179.8
25	10000063	Sub-17	15.4	183.4
19	10000184	Sub-16	15.9	181.9

Cl. 2 (6C's)



	ID	Categoría	Edad	Talla
692	10000276	Reserva	16.9	182.7
575	10000164	Sub-16	16.1	177.4
5	10000191	Sub-16	15.7	177.5

Discussion & Improvements

- Defining new reference values.
- Weighted K-means Clustering.
- PCA (Principal Component Analysis) – Dimensionality Reduction

OHIO UNIVERSITY

Thank you!

FOREVER OHIO