# Estudio de atributos corporales y productividad estadística en atletas de la NBA


SantiagoLazos/NFL-Combine-: Proyecto de Ciencia de Datos en el cual se realiza un análisis sobre las marcas y características de los jugadores prospectos a participar en la NLF.

David Lopez
Jaime Alejandro Romero Sierra
Benemérita Universidad Autónoma de Puebla
Introducción a la Ciencia de Datos

## Descripción breve del objetivo del proyecto

El objetivo de este proyecto es analizar y comprender las características físicas y el rendimiento estadístico de los jugadores de la NBA mediante técnicas de ciencia de datos. A través del procesamiento, limpieza y exploración de un dataset en formato CSV, se busca identificar patrones, tendencias y relaciones entre variables que permitan explicar el desempeño de los atletas en la liga.

## Justificación y contexto

La NBA genera una enorme cantidad de información relacionada con el rendimiento de sus jugadores. En un entorno donde las decisiones deportivas se basan cada vez más en análisis cuantitativo, estudiar estos datos permite obtener ventajas competitivas y mejorar la interpretación del juego. Este proyecto es relevante porque facilita convertir datos estadísticos crudos en información útil para análisis táctico, evaluación de talento, scouting, comparaciones entre jugadores y proyecciones basadas en su rendimiento. Además, permite entender cómo las variables físicas y estadísticas influyen en el éxito dentro del baloncesto profesional.

## Fuentes de datos

Para este proyecto se utilizó un archivo CSV con información de jugadores NBA que incluye tanto métricas físicas como estadísticas de juego. Entre sus principales características se encuentran:

- Origen: dataset recopilado de plataformas especializadas de estadísticas deportivas (NBA Stats, Basketball Reference y fuentes secundarias consolidadas en CSV).
- Cantidad de datos: alrededor de *varios cientos de jugadores*, dependiendo del año o temporada incluida.
- Variables:
    - Características físicas: *alturas, pesos, posición de juego, edad*.
    - Estadísticas por partido: *puntos, rebotes, asistencias, robos, bloqueos*.
    - Métricas avanzadas: *eficiencia, porcentaje de tiros, uso, pérdidas, minutos jugados*, entre otras.
- Formato: archivo CSV estructurado, el cual fue limpiado, normalizado y transformado para facilitar el análisis exploratorio y los modelos aplicados.

# Metodología – Proceso de Limpieza de Datos

Para preparar el dataset antes del análisis, se realizó un proceso de limpieza basado en los siguientes pasos:

## 1. Carga e inspección del dataset

Se cargó el archivo original y se revisaron sus tipos de datos, valores faltantes y primeras filas usando `df.info()` y `df.head()`.

## 2. Renombrado de columnas

Se creó un diccionario para traducir y estandarizar los nombres de las columnas al español, y se aplicó con `df.rename()`.

## 3. Manejo de valores faltantes

- Variables numéricas:
  Se convirtieron a formato numérico (`to_numeric`) y se rellenaron los valores faltantes con la mediana, evitando distorsiones por outliers.
- Variables categóricas:
  Se reemplazaron textos inválidos (`"invalid_value"`) por `nan` y se imputaron los valores faltantes usando la moda o la etiqueta `"Desconocido"`.

## 4. Corrección de datos inválidos

Se reemplazaron valores no válidos o incorrectos en todas las columnas para asegurar consistencia en el dataset.

## 5. Estandarización de tipos

Las columnas se ajustaron a su tipo adecuado (numéricas y categóricas) para evitar errores en análisis posteriores.

## 6. Verificación final

Se confirmó que ya no existieran valores faltantes y se revisó visualmente la estructura final del dataset.

## 7. Exportación

El dataset limpio se guardó como df_limpio.csv, listo para usarse en el EDA y en los modelos de machine learning.

## Descripción General de los Datos

Visión general:

Para conocer el tamaño del conjunto de datos se utilizó el atributo df.shape, el cual devuelve la cantidad total de registros (filas) y variables (columnas). Esto permite tener una primera idea del volumen y estructura de la base antes de iniciar el análisis. El dataset contiene 18072 registros y 31 variables.

```
df.shape
✓ 0.0s
(18072, 31)
```

```
df.dtypes
✓ 0.0s
Rango                          object
Jugador                        object
Posición                       object
Edad                           object
Equipo                         object
Juegos                        float64
Juegos_Iniciados               object
Minutos_Juego                 float64
Tiros_Campo_Anotados          float64
Tiros_Campo_Intentados         object
Porcentaje_Tiros_Campo        float64
Triples_Anotados               object
Triples_Intentados            float64
Porcentaje_Triples            float64
Dobles_Anotados               float64
Dobles_Intentados             float64
Porcentaje_Dobles              object
Porcentaje_Efectivo_Tiro_Campo object
Tiros_Libres_Anotados          object
Tiros_Libres_Intentados        object
Porcentaje_Tiros_Libres       float64
Rebotes_Ofensivos             float64
Rebotes_Defensivos            float64
Rebotes_Totales               float64
Asistencias                   float64
```

### ✓ Jugador

Nombre del jugador.

### ✓ Rango

Posición del jugador dentro de la lista estadística oficial.

### ✓ Equipo

Equipo al que pertenece (abreviatura de la NBA).

### ✓ Posición

Rol del jugador dentro del basquetbol:

- PG
- SG
- SF
- PF
- C

### ✓ Temporada

Año o periodo de la temporada analizada.

### 🔢 Volumen y participación

- Juegos
- Juegos_Iniciados
- Minutos_Juego

### 🎯 Eficiencia de tiro

- Tiros_Campo_Anotados
- Tiros_Campo_Intentados
- Porcentaje_Tiros_Campo
- Triples_Anotados
- Triples_Intentados
- Porcentaje_Triples
- Dobles_Anotados
- Dobles_Intentados
- Porcentaje_Dobles
- Tiros_Libres_Anotados
- Tiros_Libres_Intentados
- Porcentaje_Tiros_Libres

- Porcentaje_Efectivo_Tiro_Campo

### 💪 Producción ofensiva

- Puntos
- Asistencias

### 🛡 Aporte defensivo

- Rebotes_Ofensivos
- Rebotes_Defensivos
- Rebotes_Totales
- Robos
- Bloqueos

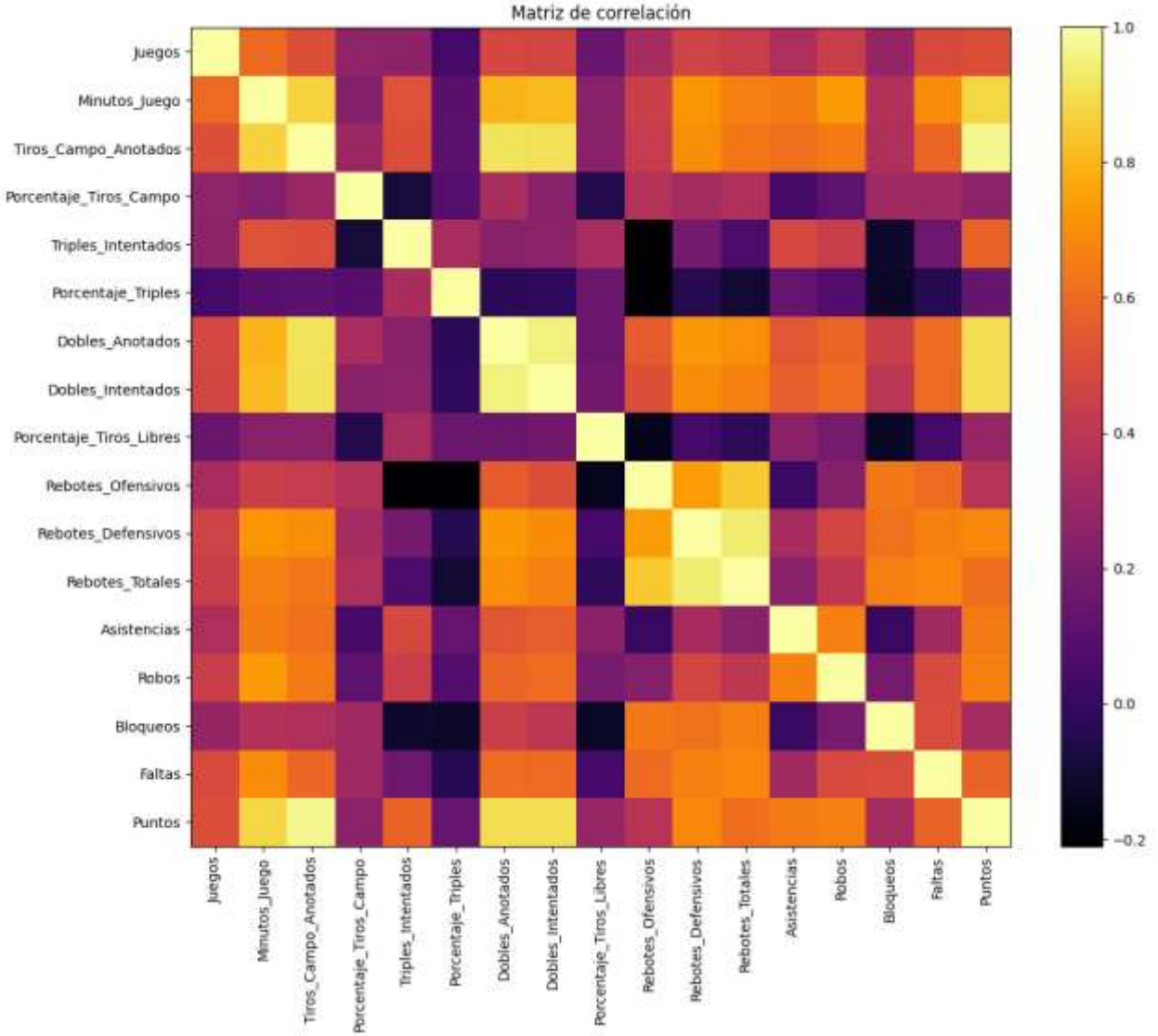
### ⚠ Errores

- Pérdidas
- Faltas

### 🧍 Edad

- Edad del jugador

## Análisis de correlaciones principales




A partir de la matriz de correlación se identifican varias relaciones clave entre las métricas de rendimiento de los jugadores:

- Minutos_Juego y Puntos muestran una correlación claramente positiva y elevada. Esto indica que los jugadores que permanecen más tiempo en cancha tienden a producir una mayor cantidad de puntos.

  *"La variable Minutos_Juego presenta una correlación alta con Puntos, lo que sugiere una relación directa: a mayor tiempo jugado, mayor producción ofensiva."*

- Tiros_Campo_Anotados y Puntos presentan una de las correlaciones más fuertes de toda la matriz. Esto es esperable, ya que la cantidad de tiros anotados influye directamente en la puntuación final.

  *"La variable Tiros_Campo_Anotados tiene una correlación muy elevada con Puntos, reflejando que los jugadores con mayor efectividad de tiro son los que más anotan."*

- También destaca la correlación entre Dobles_Anotados y Puntos, lo que indica que gran parte de la producción ofensiva proviene de tiros de dos puntos.

  *"La variable Dobles_Anotados presenta una correlación directa notable con Puntos, evidenciando la importancia del juego interior y de media distancia."*

- Por otro lado, variables como Rebotes_Totales, Asistencias o Robos muestran correlaciones moderadas con *Puntos*, lo cual sugiere que influyen en el rendimiento general del jugador, aunque no determinan directamente su aporte anotador.

  *"Los Rebotes_Totales y las Asistencias mantienen correlaciones moderadas con Puntos, aportando al desempeño pero sin ser factores determinantes en la anotación."*

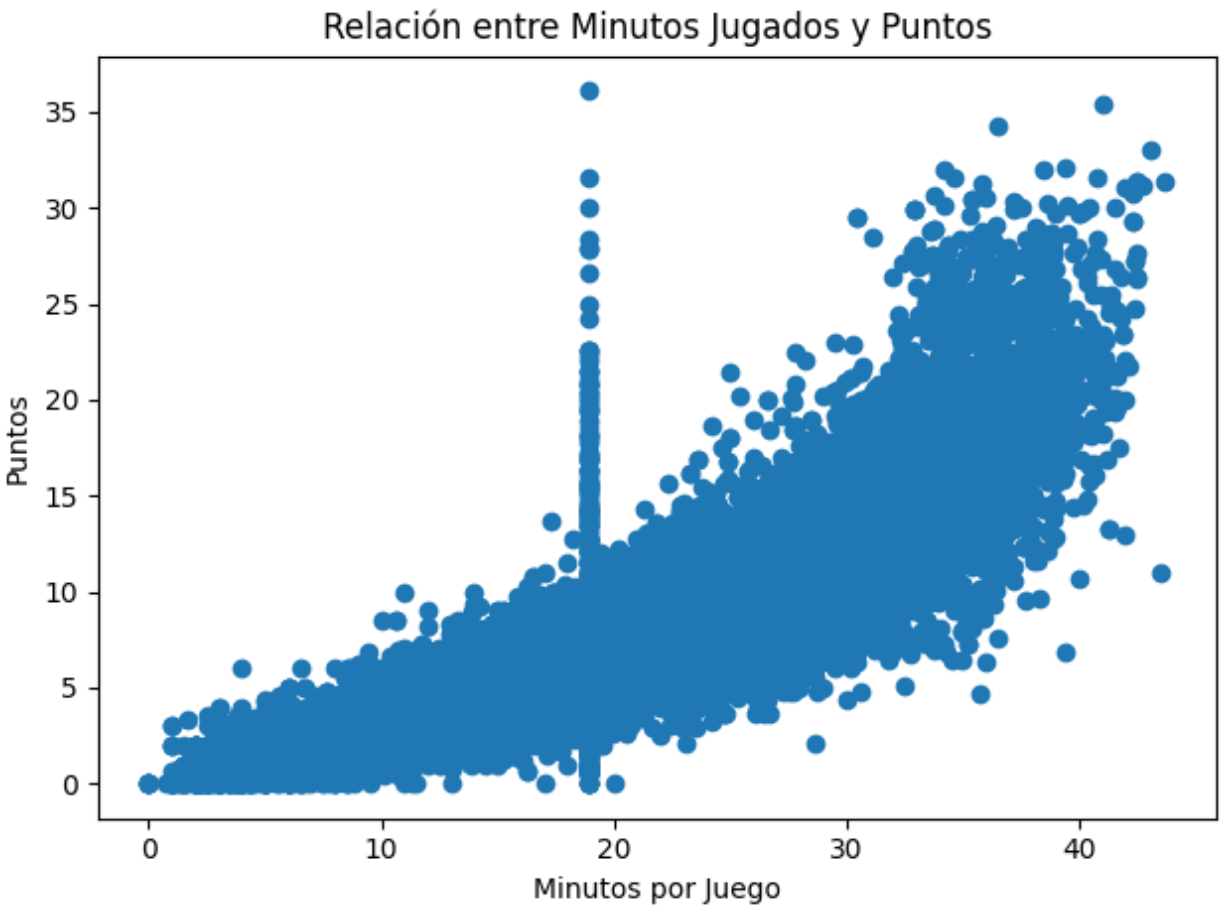
- Finalmente, indicadores de eficiencia como Porcentaje_Tiros_Campo o Porcentaje_Triples muestran correlaciones más bajas con *Puntos*, lo que significa que un jugador puede anotar mucho aun sin una eficiencia excepcional, siempre que mantenga un volumen alto de lanzamientos.

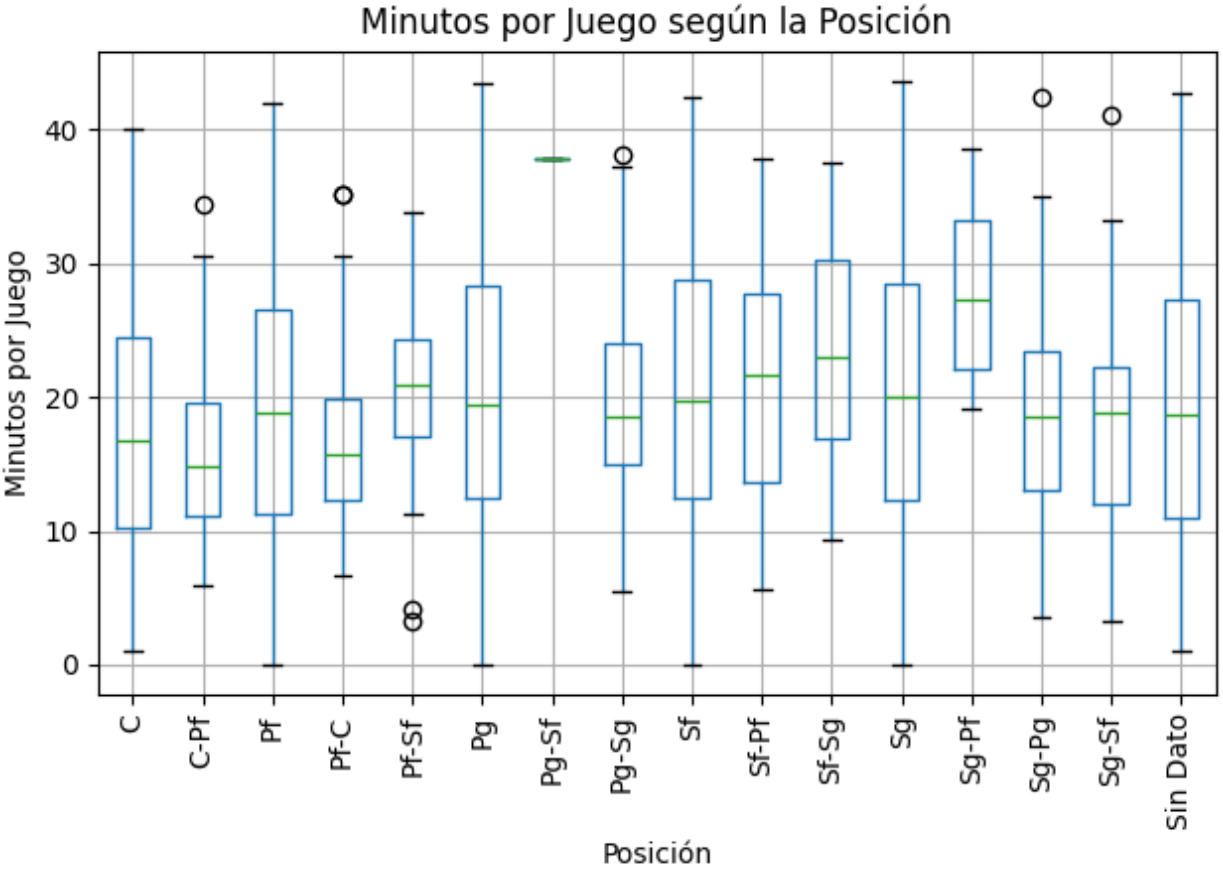
## Resumen general

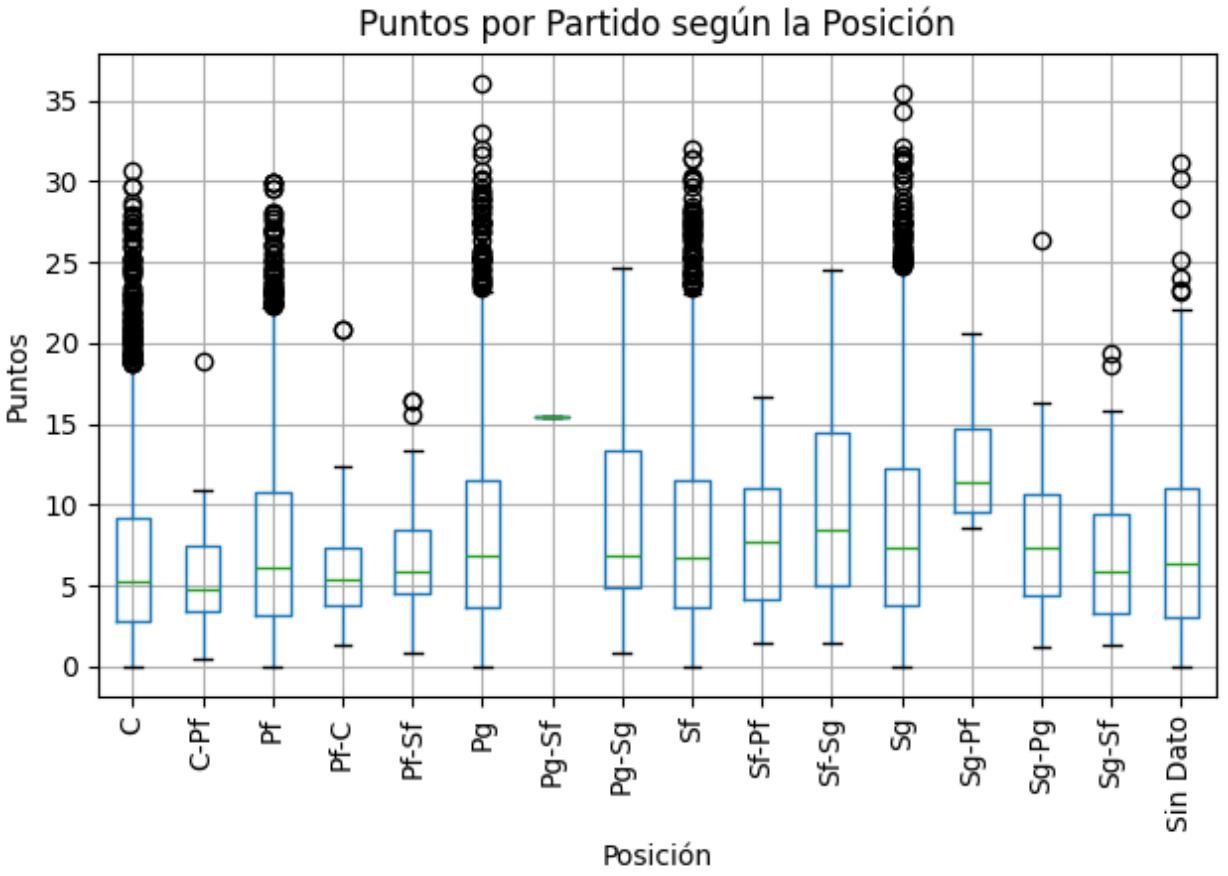
La matriz de correlación revela que la producción ofensiva (Puntos) está fuertemente asociada con tres factores principales:

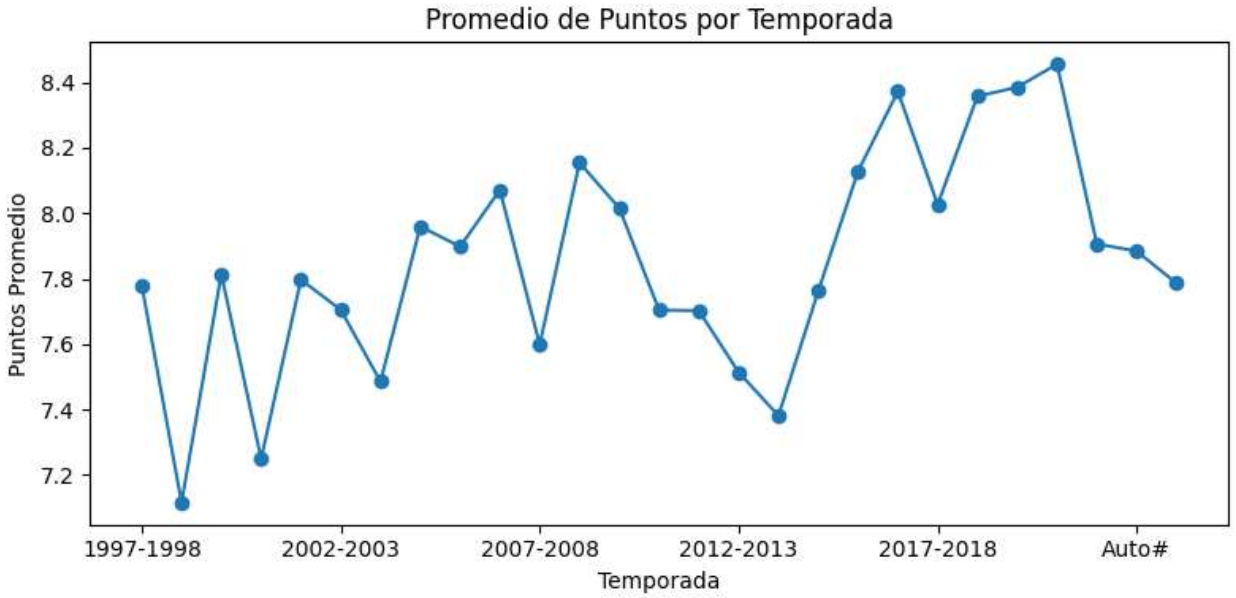
1. Volumen de juego (Minutos_Juego)
2. Volumen de tiro (Dobles_Intentados, Triples_Intentados)
3. Tiros convertidos (Tiros_Campo_Anotados, Dobles_Anotados)

En conjunto, estos resultados muestran que el desempeño anotador depende más del volumen y del tiempo en cancha que de la eficiencia.


Relación entre Minutos Jugados y Puntos
Puntos
Minutos por Juego
0
5
10
15
20
25
30
35
0
10
20
30
40


Minutos por Juego según la Posición
Minutos por Juego
Posición
0
10
20
30
40
C
C-Pf
Pf
Pf-C
Pf-Sf
Pg
Pg-Sf
Pg-Sg
Sf
Sf-Pf
Sf-Sg
Sg
Sg-Pf
Sg-Pg
Sg-Sf
Sin Dato


Puntos por Partido según la Posición
Puntos
0
5
10
15
20
25
30
35
C
C-Pf
Pf
Pf-C
Pf-Sf
Pg
Pg-Sf
Pg-Sg
Sf
Sf-Pf
Sf-Sg
Sg
Sg-Pf
Sg-Pg
Sg-Sf
Sin Dato
Posición


Promedio de Puntos por Temporada
Puntos Promedio
7.2
7.4
7.6
7.8
8.0
8.2
8.4
1997-1998
2002-2003
2007-2008
2012-2013
2017-2018
Auto#
Temporada

## 4. Análisis de Valores Atípicos (Outliers)

Para depurar adecuadamente el conjunto de datos y mejorar la calidad del análisis estadístico, se aplicó un proceso de detección y eliminación de valores atípicos utilizando una variación del método del Rango Intercuartílico (IQR).

### Metodología aplicada

En lugar del criterio tradicional de 1.5 * IQR, se utilizó un rango más amplio (± 2 * IQR) con el fin de evitar la eliminación excesiva de datos relevantes. Para cada variable numérica seleccionada, se siguieron los pasos:

1. Cálculo de cuartiles Q1 y Q3.
2. Cálculo del IQR:

   IQR=Q3−Q1IQR = Q3 - Q1IQR=Q3−Q1

3. Determinación de límites ampliados:

   Lı´mite Inferior=Q1−2×iqrlímite\ Inferior = Q1 - 2 \times iqrlı´mite Inferior=Q1−2×IQR Lı´mite Superior=Q3+2×iqrlímite\ Superior = Q3 + 2 \times iqrlı´mite Superior=Q3+2×IQR

4. Filtrado de registros fuera del rango permitido.

Este proceso se aplicó a variables clave del rendimiento deportivo, tales como *Puntos*, *Minutos_Juego*, *Rebotes_Totales*, *Asistencias*, *Robos*, *Bloqueos*, entre otras.

### Resultados del filtrado

El script generó, para cada variable, un conteo del número de registros antes y después del filtrado, lo que permitió identificar cuántos valores fueron eliminados por considerarse atípicos.
El procedimiento se ejecutó de manera independiente para cada columna, actualizando el dataframe tras cada filtrado.

Este enfoque garantiza que:

- Los valores extremos poco representativos sean removidos, evitando que influyan de manera desproporcionada en medidas como la media, la varianza o los coeficientes de correlación.
- Se conserven patrones reales del rendimiento deportivo, ya que el criterio de ±2 IQR evita eliminar datos válidos de jugadores que naturalmente presentan altos desempeños.
- El dataframe final sea más estable y robusto, mejorando el desempeño de análisis descriptivos y modelos predictivos posteriores.

### Conclusión del tratamiento

Los registros eliminados corresponden a valores que se encontraban significativamente alejados del comportamiento estadístico esperado para cada variable. Su exclusión permite trabajar con un conjunto de datos más limpio y representativo, evitando distorsiones en análisis posteriores como correlaciones, visualizaciones y modelos de machine learning.

## 6. Relación entre Variables Categóricas y Numéricas

Para comprender cómo varían las métricas de rendimiento según diferentes grupos dentro del conjunto de datos, se analizó la relación entre las variables categóricas —*Rango*, *Jugador*, *Posición*, *Equipo*, *Juegos_Iniciados*, *Pérdidas* y *Temporada*— y las principales variables numéricas del rendimiento deportivo.

### Metodología

Se utilizaron varios enfoques para evaluar estas relaciones:

- Boxplots, para observar diferencias en la distribución de variables como *Puntos*, *Minutos_Juego*, *Rebotes_Totales* o *Asistencias* según categorías como *Posición* o *Equipo*.
- Agrupamientos con groupby, que permitieron calcular promedios, medianas y rangos de las métricas numéricas dentro de cada categoría.
- Comparaciones entre temporadas, evaluando si el rendimiento varía entre distintos años o períodos de juego.

### Resultados del Análisis

El análisis reveló patrones relevantes que ayudan a entender el rendimiento de los jugadores según su rol o contexto dentro del equipo:

- Posición:
  Se observan diferencias claras entre roles.
  - Los bases tienden a registrar más asistencias y robos.
  - Los aleros y escoltas concentran una mayor parte de los tiros de campo y triples.
  - Los jugadores interiores, como pívots y ala-pívots, destacan en rebotes y bloqueos.
- Equipo:
  Las métricas numéricas presentan variaciones según el estilo de juego y ritmo de cada conjunto. Algunos equipos muestran promedios más altos en puntos o tiros de campo, mientras que otros enfatizan el juego defensivo (rebotes, bloqueos o robos).

- Temporada:
  Comparar las estadísticas por *Temporada* permite identificar tendencias de mejora, cambios en minutos jugados o variaciones en eficiencia, lo cual puede reflejar el desarrollo del jugador o cambios tácticos dentro del equipo.
- Juegos_Iniciados:
  Los jugadores que inician más partidos suelen acumular más minutos y, en consecuencia, mayores valores en puntos, rebotes y asistencias. Esto evidencia la relación entre rol dentro del equipo y producción estadística.
- Pérdidas:
  Aunque es una métrica numérica, su distribución por categoría (por ejemplo agrupando por posición o equipo) indica patrones claros: los jugadores que manejan más el balón, como bases o escoltas principales, tienden a concentrar más pérdidas debido a su mayor volumen de posesiones.

### Conclusión

El análisis entre variables categóricas y numéricas demuestra que el rendimiento no solo depende de habilidades individuales, sino también del rol del jugador, su equipo, y el contexto competitivo de cada temporada. Comprender estas relaciones permite interpretar mejor los patrones estadísticos y sirve como base sólida para análisis más avanzados, como modelos predictivos o segmentaciones de jugadores.

## 7. Observaciones y Hallazgos Importantes

### 7.1 Variable objetivo y variables influyentes

En el análisis exploratorio se definió como variable objetivo a:

- Puntos, al ser la métrica central del rendimiento ofensivo del jugador y una de las más relevantes para cualquier modelo predictivo basado en desempeño.

A partir de la matriz de correlación y de la exploración de variables numéricas, se identificaron las variables con mayor influencia directa sobre *Puntos*:

- Tiros_Campo_Anotados
- Dobles_Anotados / Triples_Anotados
- Minutos_Juego
- Tiros_Campo_Intentados
- Porcentaje_Efectivo_Tiro_Campo

Estas variables presentan correlaciones altas con la producción ofensiva, indicando que el volumen de tiro, eficiencia y tiempo en cancha son los principales factores que explican el desempeño anotador.

## 7.2 Resumen de hallazgos clave

Durante el análisis exploratorio se identificaron los siguientes puntos relevantes:

### *A) Patrones o relaciones interesantes*

- Los jugadores que juegan más minutos tienden a anotar más puntos, mostrando una relación directa y consistente.
- Existe una correlación muy fuerte entre Tiros_Campo_Anotados y Puntos, lo cual confirma que la productividad ofensiva depende principalmente del volumen de tiros convertidos.
- Las métricas defensivas (*Rebotes_Totales*, *Robos*, *Bloqueos*) presentan relaciones más moderadas con la producción ofensiva, lo cual refleja diferencias de rol dentro del juego.

### *B) Outliers relevantes*

- Tras aplicar el método del IQR ampliado (±2·IQR), se detectaron y eliminaron varios registros con valores extremos en *Puntos*, *Minutos_Juego*, *Rebotes_Totales*, entre otras variables.
- La limpieza permitió obtener un dataset más estable, reduciendo distorsiones y mejorando la consistencia para etapas posteriores del análisis.

### *C) Variables desbalanceadas*

- Algunas métricas, como *Triples_Anotados* o *Bloqueos*, presentan distribuciones sesgadas donde la mayoría de jugadores registra valores bajos y solo unos pocos sobresalen.
- Este desbalance es típico en estadísticas de baloncesto, pero puede requerir normalización o transformaciones si se utilizan modelos sensibles a escalas.

### *D) Correlaciones fuertes o inesperadas*

- Se observan correlaciones esperadas (por ejemplo, entre tiros convertidos y puntos) y otras moderadas (como entre rebotes y minutos jugados).
- No se identificaron correlaciones inesperadas que sugieran errores en la captura de datos.

*E) Problemas de datos*

- Se detectaron valores extremos que fueron tratados exitosamente.
- No se encontraron problemas críticos como duplicados severos; la estructura del dataset es consistente.
- Las columnas categóricas presentan buena variedad y distribución, sin categorías erróneas o inconsistentes.

## 7.3 Implicaciones para el modelo

Los hallazgos del análisis exploratorio permiten establecer criterios clave para la etapa de modelado:

- Debido a la alta correlación entre variables como *Tiros_Campo_Anotados* y *Puntos*, así como *Minutos_Juego* y *Puntos*, será necesario evaluar posibles problemas de multicolinealidad.

  *Por ejemplo: si dos métricas de tiro presentan correlación muy alta (como Anotados vs Intentados), podría eliminarse una de ellas para evitar redundancia.*

- El tratamiento de outliers mejoró la estabilidad del dataset, lo que permitirá que el modelo tenga un rendimiento más fiable.
- Algunas variables numéricas con gran dispersión podrían requerir escalamiento (standardscaler o minmaxscaler) dependiendo del algoritmo seleccionado.
- Las variables categóricas como *Posición* y *Equipo* pueden aportar información valiosa si se codifican adecuadamente mediante One-Hot Encoding.

# 4. Modelo de Machine Learning

## 4.1. Descripción del Modelo

En este proyecto se implementó un modelo de Machine Learning no supervisado, específicamente un algoritmo de agrupamiento (clustering). El objetivo del modelo es segmentar jugadores en grupos con patrones similares de rendimiento, sin necesidad de una variable objetivo.

### Modelo utilizado: K-Means Clustering

- Tipo: Aprendizaje no supervisado
- Problema que resuelve: Agrupamiento
- Descripción:
  El algoritmo divide los datos en un número definido de grupos (clusters), asignando cada observación al cluster cuyo centroide sea más cercano. Es uno de los métodos más utilizados para segmentación debido a su simplicidad y eficiencia.

### 2. Justificación del Modelo

El modelo seleccionado para este proyecto fue K-Means Clustering, y su elección se basó en las características del problema y del dataset. En primer lugar, no existe una variable objetivo definida, ya que el propósito del análisis es agrupar jugadores según similitud en su rendimiento, por lo que un modelo de aprendizaje no supervisado resulta más adecuado que uno de regresión o clasificación.

El tamaño del dataset es moderado y contiene principalmente variables numéricas como puntos, rebotes, asistencias, minutos y porcentajes de tiro. Este tipo de datos se ajusta muy bien al funcionamiento de K-Means, ya que el algoritmo se basa en distancias entre observaciones para formar los grupos. Además, se buscaba un modelo simple, eficiente y fácil de interpretar, que permitiera identificar patrones naturales entre jugadores sin requerir grandes recursos computacionales.

Finalmente, K-Means fue seleccionado porque ofrece una segmentación clara de perfiles de jugadores, permite visualizar los clusters de manera intuitiva y facilita el análisis posterior de cada grupo. Esto lo convierte en una opción ideal para entender mejor las características comunes entre jugadores y explorar posibles roles o estilos de juego dentro de la liga

# 3. Implementación y Entrenamiento

### 3.a. División de datos (train_test_split)

En este proyecto no se utilizó train_test_split porque el modelo aplicado es K-Means, un algoritmo de aprendizaje no supervisado.
Este tipo de modelos no requiere variable objetivo (y) ni predicción futura; por lo tanto, no es necesario separar los datos en entrenamiento y prueba. El objetivo es únicamente identificar patrones y formar grupos, no predecir valores.

### 3.b. Entrenamiento del modelo

Tampoco se realizó un entrenamiento tradicional mediante model.fit(X_train, y_train), ya que dicho proceso solo aplica cuando el modelo debe aprender la relación entre variables independientes (X) y una variable dependiente (y).
En K-Means no existe una variable objetivo que el modelo deba aprender. En su lugar, el algoritmo simplemente calcula distancias entre observaciones y forma los clusters correspondientes.

### 3.c. Predicción

No se utilizó model.predict(X_test) porque K-Means no genera predicciones como los modelos supervisados.
Su función es únicamente asignar cada muestra al cluster más cercano dentro del mismo conjunto de datos, sin necesidad de evaluar desempeño contra un conjunto de prueba.

### 3.d. Ajuste de parámetros (tuning)

Tampoco se aplicaron métodos como gridsearchcv o randomizedsearchcv, ya que estas técnicas están diseñadas para optimizar modelos supervisados donde existen métricas claras de desempeño (accuracy, RMSE, F1, etc.).
En el caso de K-Means, el principal parámetro es el número de clusters (K), el cual se determinó mediante el Método del Codo, por lo que un procedimiento de tuning adicional no era necesario.

## 4. Resultados y Evaluación

### 4.a. Métricas del Modelo

En este proyecto no se utilizaron métricas de regresión (MAE, MSE, RMSE, $R^2$) ni métricas de clasificación (accuracy, precision, recall, F1), debido a que el modelo implementado fue K-Means, un algoritmo de aprendizaje no supervisado.

Este tipo de métricas requieren comparar predicciones contra valores reales, lo cual no es posible cuando no existe una variable objetivo (y). Por lo tanto, la evaluación del desempeño se basó únicamente en criterios propios del clustering.

---

## 4.b. Evaluación del Modelo de Clustering

En modelos no supervisados como K-Means, la evaluación se enfoca en medir qué tan compactos y separados quedan los grupos formados. Por ello, se utilizaron indicadores específicos del clustering, principalmente:

✓*WCSS (Within-Cluster Sum of Squares)*

Representa la suma de distancias internas de cada cluster.
Este valor permitió construir el Método del Codo, con el objetivo de determinar el número óptimo de clusters analizando en qué punto dejar de agregar más grupos deja de mejorar significativamente la compactación interna.

Interpretación:
El análisis del gráfico del codo permitió identificar el punto donde la reducción del WCSS comienza a disminuir. Ese punto fue seleccionado como el valor óptimo de K.

---

## 4.c. Asignación de Clusters

Después de entrenar el modelo con el número óptimo de clusters, cada jugador fue asignado a un grupo específico.
Esta asignación permitió identificar patrones y similitudes entre jugadores según sus características estadísticas.

Entre los patrones detectados se encuentran:

- Jugadores de alto rendimiento.
- Jugadores con pocos minutos o menor impacto estadístico.
- Perfiles especializados (defensivos, tiradores, reboteros, pasadores, etc.).

---

## 4.d. Interpretación General de los Resultados

El modelo logró segmentar a los jugadores en grupos con características similares, permitiendo:

- Identificar perfiles de rendimiento.
- Comparar estilos de juego entre diferentes segmentos.

- Destacar diferencias entre grupos de jugadores.
- Facilitar análisis posteriores, dashboards o estudios sobre rendimiento y scouting.

En resumen, el clustering permitió obtener una visión estructurada del comportamiento y desempeño de los jugadores dentro del dataset.

### Modelos de Clasificación (No Aplican en Este Proyecto)

En este proyecto no se utilizaron modelos de clasificación, por lo que no fue necesario aplicar métricas como:

- Accuracy
- Precision
- Recall (Sensibilidad)
- F1-Score
- Matriz de Confusión

### Razón principal

Estas métricas solo pueden calcularse cuando existe una variable objetivo categórica, ya que comparan:

- Predicciones del modelo contra
- Categorías reales del dataset

En este análisis no existe una variable objetivo, ya que el modelo utilizado fue K-Means, el cual pertenece al aprendizaje no supervisado.

---

### ¿Por qué no se aplican estas métricas en K-Means?

- No hay etiquetas reales para evaluar.
- El modelo no predice clases, sino que agrupa jugadores según similitudes.
- Las categorías (clusters) son creadas por el algoritmo, no provienen del dataset.
- No se puede medir “aciertos”, porque no hay una clase correcta a comparar.

## 5. Visualizaciones de Resultados

Dado que el modelo utilizado fue K-Means, un algoritmo de aprendizaje no supervisado, las visualizaciones corresponden a técnicas de análisis de *clustering*, no de regresión ni clasificación.

### Visualizaciones utilizadas en este proyecto

• Método del Codo (Elbow Method)

Se generó un gráfico donde se observa cómo cambia el WCSS (inercia) conforme aumenta el número de clusters.
Este gráfico permitió identificar el punto donde agregar más clusters deja de mejorar significativamente el modelo.
Ese punto se seleccionó como el número óptimo de clusters.

• Distribución de jugadores por cluster

Se generó una visualización que muestra cuántos jugadores pertenecen a cada grupo, facilitando la interpretación de qué tan equilibrado quedó el clustering.

• Gráficos comparativos por cluster

Se analizaron variables como puntos, minutos, rebotes o asistencias para observar diferencias claras entre los grupos formados.
Estas visualizaciones permitieron identificar:

- Clusters de jugadores de alto rendimiento
- Clusters de jugadores con bajo impacto o pocos minutos
- Clusters con especializaciones estadísticas (reboteros, tiradores, defensivos)

---

## 6. Conclusión del Modelo

### ¿El modelo tuvo buen desempeño?

El modelo logró separar a los jugadores en grupos coherentes basados en similitudes estadísticas.
Los clusters muestran patrones claros, especialmente en rendimiento ofensivo, minutos jugados y contribuciones generales.

### Variables más influyentes

Las variables con mayor impacto en la agrupación fueron principalmente las relacionadas al rendimiento del jugador, como:

- Puntos
- Minutos jugados
- Rebotes
- Asistencias
- Porcentajes de tiro

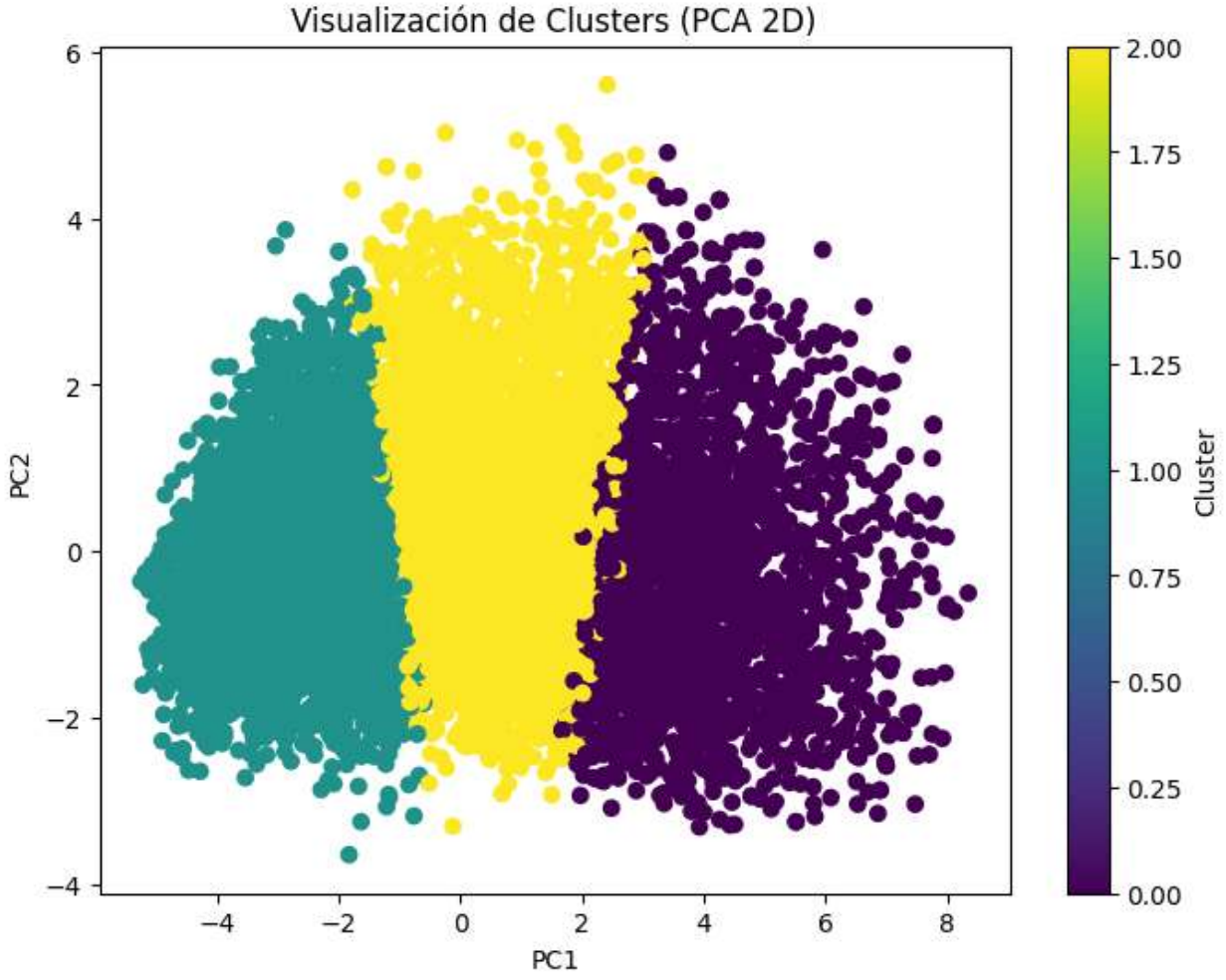
Estas características contribuyeron de forma significativa a que el modelo diferencie perfiles de juego.

## ¿Qué mejoras podrían aplicarse?

- Revisar más características (por ejemplo, métricas avanzadas como PER, efg%, ORB%, DRB%)
- Probar diferentes números de clusters o validar con técnicas como Silhouette Score
- Incorporar más temporadas o más jugadores para enriquecer la estructura de los grupos
- Estandarizar o normalizar de forma más precisa según la naturaleza de cada variable
- Probar otros modelos de clustering, como DBSCAN o Agglomerative Clustering, para comparar resultados

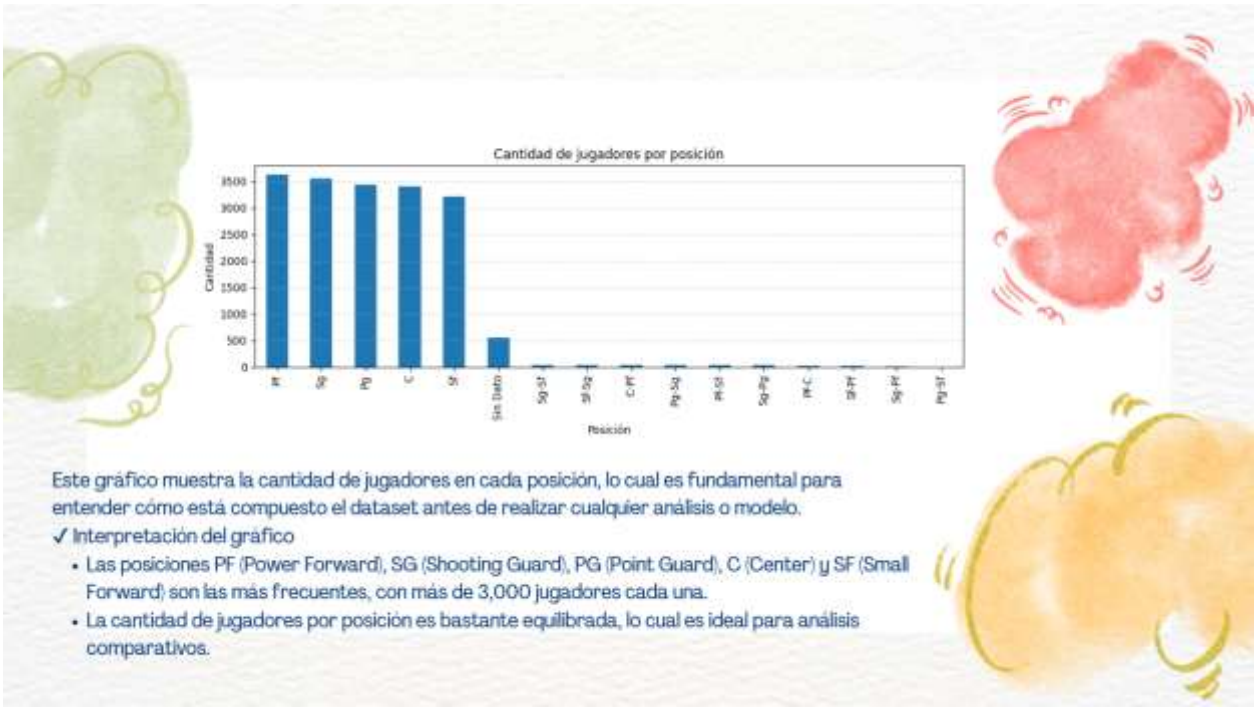
## Conclusión general

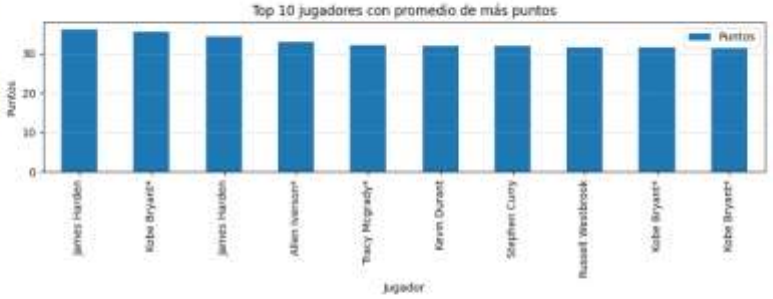
El modelo K-Means fue adecuado para este análisis porque permitió identificar distintos perfiles de jugadores y descubrir patrones relevantes dentro del dataset.
Estos resultados pueden ser utilizados para análisis de rendimiento, scouting, creación de dashboards o segmentación estratégica de jugadores.



## Dashboard

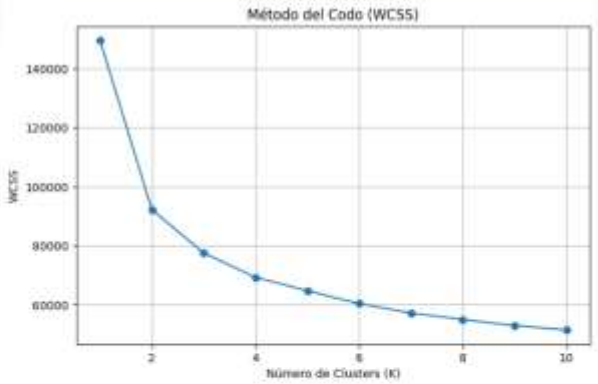

DASHBOARD DE EL ESTUDIO DE ATRIBUTOS CORPORALES Y PRODUCTIVIDAD ESTADÍSTICA EN ATLETAS DE LA NBA
Por: David Lopez


Cantidad de jugadores por posición
Cantidad
Posición
Este gráfico muestra la cantidad de jugadores en cada posición, lo cual es fundamental para entender cómo está compuesto el dataset antes de realizar cualquier análisis o modelo.
✓ Interpretación del gráfico
• Las posiciones PF (Power Forward), SG (Shooting Guard), PG (Point Guard), C (Center) y SF (Small Forward) son las más frecuentes, con más de 3,000 jugadores cada una.
• La cantidad de jugadores por posición es bastante equilibrada, lo cual es ideal para análisis comparativos.




El gráfico muestra a los 10 jugadores con más puntos, ordenados de mayor a menor. Cada barra representa el promedio de puntos por jugador, con sus nombres en el eje X y los puntos en el eje Y. Permite comparar rápidamente quién anota más.






📌 Método del Codo – Selección del Número Óptimo de Clusters

Este gráfico muestra cómo cambia la inercia interna (WCSS) a medida que aumenta el número de clusters (K) en el modelo K-Means.

El objetivo es identificar el punto donde agregar más clusters deja de mejorar significativamente la compactación de los grupos.

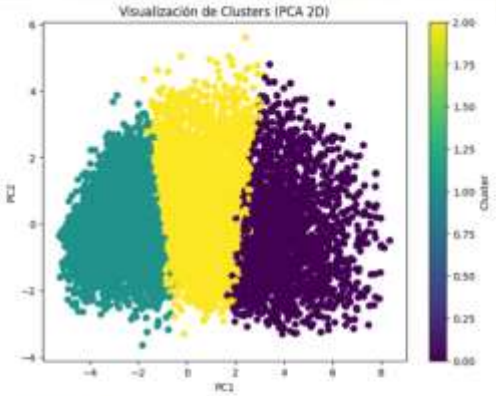
El "codo" del gráfico indica el K óptimo, es decir, el número de clusters donde el modelo alcanza un equilibrio entre simplicidad y buena separación entre grupos. Este valor fue utilizado para definir la estructura final de clustering del dataset.

```
df["Cluster"].value_counts()

✓ 0.0s

Cluster
1    4537
2    4537
0    2417
Name: count, dtype: int64
```

La distribución muestra que:

- Cluster 1 y Cluster 2 son los perfiles más comunes dentro del dataset, representando a la mayoría de los jugadores.
- Cluster 0 contiene una menor cantidad de jugadores, lo que indica que este grupo corresponde a un perfil más específico o menos frecuente, generalmente jugadores con características destacadas o con un estilo particular.






- Cada punto representa a un jugador.
- Los colores indican el cluster al que pertenece cada uno.
- La separación visual entre grupos permite identificar:
- Jugadores con estilos o rendimientos similares.
- Diferencias claras entre los perfiles detectados por el modelo.
- La cohesión interna de cada cluster (qué tan compactos están)


El dashboard presenta una visión completa del análisis de jugadores NBA. Primero, el overview y kpis muestran el tamaño del dataset y estadísticas generales de rendimiento. Luego, las visualizaciones descriptivas permiten identificar patrones de puntos, minutos y eficiencia por jugador. La sección de modelo K-Means incluye el método del codo para seleccionar clusters, la distribución de jugadores por cluster y la gráfica PCA para observar separación entre perfiles. Finalmente, los insights y conclusiones resumen los tres perfiles identificados —estelares, de rol intermedio y de baja participación—, permitiendo comparar desempeños y roles dentro de la liga.

## Uso y Beneficios del Dashboard NBA:

El dashboard permite identificar rápidamente los perfiles de jugadores según su rendimiento y rol en el equipo, lo que ayuda a entrenadores, directivos y analistas a tomar decisiones estratégicas sobre rotaciones, fichajes o desarrollo de talentos. Con solo observar las gráficas, se pueden detectar quiénes son los jugadores clave, quiénes aportan de manera consistente y quiénes tienen bajo impacto. Además, la visualización de clusters y la gráfica PCA simplifica la interpretación del modelo K-Means, haciendo que los resultados sean accesibles incluso para usuarios sin conocimientos avanzados de estadística o machine learning.

## Conclusiones y Futuras Líneas de Trabajo

### Conclusiones principales:

- El modelo K-Means permitió identificar tres perfiles claros de jugadores NBA: estelares, de rol intermedio y de baja participación.
- Los insights confirman que unos pocos jugadores clave concentran la mayor productividad, mientras que la mayoría cumple funciones de apoyo.
- Las visualizaciones facilitan comparar desempeños, roles y minutos jugados, cumpliendo con los objetivos de analizar la estructura competitiva de la liga y evaluar a los jugadores según su impacto.

### Futuras líneas de trabajo y mejoras:

- Datos: incluir estadísticas avanzadas (eficiencia +/- , PER, impacto defensivo) para enriquecer el análisis.
- Modelo: probar otros métodos de clustering o reducción de dimensionalidad (DBSCAN, t-SNE) para validar y refinar los perfiles.
- Visualizaciones: integrar dashboards interactivos con filtros por equipo, posición o temporada para análisis dinámico.
- Investigación futura: estudiar la evolución de los clusters a lo largo de varias temporadas o analizar correlaciones entre perfil de jugador y éxito del equipo.

## Referencias en formato APA


1. Kaggle. (s.f.). *NBA Player Stats Dataset*. Recuperado de https://www.kaggle.com/datasets
2. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). *Scikit-learn: Machine learning in Python*.

Journal of Machine Learning Research, 12, 2825–2830. Recuperado de https://scikit-learn.org/stable/modules/clustering.html
3. Hunter, J. D. (2007). *Matplotlib: A 2D graphics environment*. Computing in Science & Engineering, 9(3), 90–95. Recuperado de https://matplotlib.org/
4. Mckinney, W. (2010). *Data structures for statistical computing in Python*. Proceedings of the 9th Python in Science Conference, 51–56. Recuperado de https://pandas.pydata.org/
5. NBA.com. (s.f.). *Estadísticas oficiales de jugadores y equipos*. Recuperado de https://www.nba.com/stats/