

|  |
| --- |
| MAESTRÍA Y ESPECIALIZACIÓN EN MÉTODOS CUANTITATIVOS PARA LA GESTIÓN Y ANÁLISIS DE DATOS EN ORGANIZACIONES |
| Clasificación y Agrupación de Jugadores de la NBA para la temporada 2024: Un Análisis Multivariante del Rendimiento |
| Métodos de Análisis Multivariante |

|  |
| --- |
| Autor: Jesús Ochoa  Profesores: Dra. Blaca Vitale, Act. Pablo Caviziel, Act. Miguel Cordoba |

Contenido

[Resumen 2](#_Toc170320016)

[Introducción 3](#_Toc170320017)

[Planteamiento del Problema 4](#_Toc170320018)

[Objetivo 4](#_Toc170320019)

[Alcance 5](#_Toc170320020)

[Hipótesis 6](#_Toc170320021)

[Marco Teórico 7](#_Toc170320022)

[Antecedentes de la investigación 7](#_Toc170320023)

[Bases Teóricas 8](#_Toc170320024)

[Obtención de Datos 12](#_Toc170320025)

[Desarrollo Análisis y Resultados 13](#_Toc170320026)

[Análisis Descriptivo 14](#_Toc170320027)

[Modelo Manova 21](#_Toc170320028)

[Análisis de Componentes Principales 23](#_Toc170320029)

[PCA 25](#_Toc170320030)

[Clustering 28](#_Toc170320031)

[Análisis Discriminante 35](#_Toc170320032)

[Conclusiones 41](#_Toc170320033)

[Bibliografía 43](#_Toc170320034)

## Resumen

El objetivo de este estudio es crear un modelo multivariante utilizando métricas clave de rendimiento para clasificar y agrupar a los jugadores de la NBA para la temporada 2024. Se utilizará un análisis multivariante de la varianza (MANOVA) utilizando un enfoque estructurado para evaluar las diferencias significativas en el Impacto del Jugador Estimado (PIE) y el Porcentaje de Uso (USG%) en relación con la posición y la edad del jugador. Posteriormente, se utilizará un análisis de componentes principales (PCA) para reducir la dimensionalidad de los datos y crear una variable compuesta llamada “Rendimiento”, que representa el rendimiento general de los jugadores. Se utilizará esta nueva variable para agrupar a los jugadores utilizando análisis de clusters para identificar grupos homogéneos de rendimiento. Finalmente, se implementará un análisis discriminante para clasificar a los jugadores en sus respectivas categorías de rendimiento. Esto evaluará la precisión y la capacidad predictiva del modelo creado. Este enfoque completo facilitará una comprensión profunda y práctica del rendimiento de los jugadores, lo que facilitará la clasificación y agrupación de los jugadores en función de una variedad de métricas de rendimiento. El estudio concluyó que, si bien el modelo identificó patrones distintivos de rendimiento, el análisis discriminante mostró limitaciones en la precisión de la clasificación. Este hallazgo sugiere la necesidad de refinar las variables utilizadas o considerar enfoques alternativos para mejorar la precisión del modelo.

Además, la base de datos y los códigos para replicar este informe interactivo están disponibles en el siguiente repositorio de GitHub: [Analisis-Multivariante-NBA](https://github.com/jarof13/Analisis-Multivariante-NBA.git).

## Introducción

Tanto para los analistas deportivos como para los equipos profesionales que buscan optimizar sus alineaciones y estrategias, el análisis del rendimiento de los jugadores de baloncesto en la NBA ha sido un tema de creciente interés. La Liga Nacional de Baloncesto (NBA) ha experimentado un cambio significativo desde su fundación en 1946 hasta convertirse en la liga de baloncesto más prestigiosa y competitiva del mundo. La capacidad de evaluar y comparar el rendimiento de los jugadores de manera precisa y objetiva es crucial para la toma de decisiones dentro de los equipos, ya sea para planificar entrenamientos, crear estrategias de juego o contratar nuevos talentos.

La necesidad de desarrollar herramientas analíticas sofisticadas que puedan sintetizar y simplificar la gran cantidad de datos recolectados durante cada temporada es el fundamento de la elección de este tema. La cantidad de datos disponibles ha aumentado exponencialmente desde la introducción de tecnologías de seguimiento y estadísticas avanzadas. Aplicar métodos multivariantes, como el análisis de componentes principales (PCA), el análisis de clusters y el análisis discriminante, se convierte en una estrategia efectiva para reducir la dimensionalidad de los datos y extraer información valiosa en este contexto.

Geopolíticamente, la NBA tiene un impacto global, con jugadores internacionales y una audiencia que se extiende por todo el mundo. También tiene un impacto en Estados Unidos, su país de origen. El vínculo entre este estudio y el entorno se ve en cómo los datos de rendimiento de los jugadores pueden afectar las decisiones financieras y estratégicas de las franquicias, que afectan no solo el juego en sí, sino también aspectos comerciales y de marketing. Este análisis también ayuda a desarrollar técnicas estadísticas que pueden usarse en otros deportes y áreas donde la evaluación del rendimiento es importante. El creciente impacto de la analítica y la inteligencia artificial se puede ver en la integración de técnicas avanzadas de análisis de datos en el deporte.

# Planteamiento del Problema

En la presente sección se expone formalmente la delimitación en el cual se desarrollará la investigación estadística dada la importancia de clasificar y agrupar a los jugadores de la NBA según su rendimiento, partiendo de las diferentes métricas ya existentes.

## Objetivo

El objetivo principal de este estudio es desarrollar y abordar el problema específico de evaluar y clasificar el rendimiento de los jugadores de la National Basketball Association de manera precisa y eficiente mediante el uso de técnicas estadísticas multivariantes. Se presenta un desafío identificar las métricas más relevantes y cómo estas interactúan para influir en el rendimiento global de un jugador en el contexto de una liga tan competitiva y con un volumen de datos tan amplio y diverso.

Para lograr este objetivo, se plantea la siguiente problemática específica: ¿Cómo se pueden utilizar métodos multivariantes para identificar diferencias significativas en las métricas clave de rendimiento de los jugadores de la NBA, reducir la complejidad de los datos, agrupar jugadores en categorías homogéneas y, finalmente, clasificar con precisión a los jugadores según su rendimiento?

Esta problemática se desglosa en varios subobjetivos:

1. Evaluar las diferencias significativas con metricas existentes: Aplicar un análisis multivariante de la varianza (MANOVA) para determinar si las métricas clave como el Impacto del Jugador Estimado (PIE) y el Porcentaje de Uso (USG%) varían significativamente según la posición del jugador y su grupo de edad.
2. Reducir la dimensionalidad: Utilizar el análisis de componentes principales (PCA) para identificar las métricas más importantes que explican la mayor parte de la variabilidad de las más de 60 variables que se tienen en la base de datos.
3. Crear una variable compuesta de rendimiento: Construir una nueva variable llamada “Raiting”, basada en las componentes principales, que represente el rendimiento global de los jugadores.
4. Agrupar jugadores: Emplear análisis de clusters para agrupar a los jugadores en grupos homogéneos basados en los componentes principales, la variable “Raiting”, la edad y la posición.
5. Clasificar y validar: Implementar un análisis discriminante para clasificar a los jugadores en sus respectivas categorías de rendimiento y evaluar la capacidad predictiva del modelo desarrollado.

## Alcance

El alcance de este trabajo se delimita claramente para asegurar la validez de las conclusiones obtenidas a partir del análisis multivariante del rendimiento de los jugadores de la NBA. El estudio abarca las siguientes áreas:

1. Período de Estudio: El análisis se centra en los datos de las estadísticas de los jugadores para el año 2024.
2. Población de Estudio: Se incluyen todos los jugadores que han participado en la temporada del período mencionado. Además, que hayan participado en un porcentaje de juegos disputados mayor al 50% de los juegos de la temporada y que jugaran al menos una cantidad de minutos superior a 8 minutos de juegos en estos partidos, esto para poder tener unos datos menos dispersos y más acorde a una situación de un jugador realmente activo en un equipo. Esto garantiza que el análisis refleje una amplia variedad de estilos de juego y niveles de rendimiento.
3. Métricas de Rendimiento: Se consideran múltiples métricas clave de rendimiento, incluyendo pero no limitándose a: puntos anotados (PTS), rebotes (REB), asistencias (AST), robos (STL), bloqueos (BLK), pérdidas (TOV), porcentaje de tiros de campo (FG%), porcentaje de tiros de tres puntos (3P%), porcentaje de tiros libres (FT%), rebotes ofensivos (OREB), rebotes defensivos (DREB), posesiones (POSS), Impacto del Jugador Estimado (PIE), Porcentaje de Uso (USG%), Edad (Age) y posición en la que juega el jugador (Position).

### Limitaciones

Las conclusiones están delimitadas a los jugadores y una temporada especifica, por ende, pueden no ser directamente extrapolables a otros contextos o ligas sin un análisis adicional.

## Hipótesis

En el contexto del análisis multivariante del rendimiento de los jugadores en la temporada 2024, se plantean las siguientes hipótesis que deberán ser contrastadas a lo largo del estudio:

* H1: Existen diferencias significativas en las métricas de rendimiento, específicamente el Impacto del Jugador Estimado (PIE) y el Porcentaje de Uso (USG%), según la posición del jugador y su grupo de edad.
* H2: La reducción de dimensionalidad mediante el análisis de componentes principales (PCA) permitirá identificar un conjunto reducido de componentes que explican la mayor parte de la variabilidad de las métricas de los jugadores.
* H3: La nueva variable “Raiting”, construida a partir de los componentes principales obtenidos del PCA, será una representación válida y robusta del rendimiento global de los jugadores.
* H4: Los jugadores pueden ser agrupados en clusters homogéneos en función de la variable “Raiting”, su edad, su posición de juego y los componentes principales, identificando grupos de rendimiento similar.
* H5: Un modelo de análisis discriminante, basado en las componentes principales y los clusters, podrá clasificar con precisión a los jugadores en sus respectivas categorías de rendimiento y edad, demostrando una alta capacidad predictiva.

# Marco Teórico

En este apartado se presentarán los antecedentes de las investigaciones que sirvieron como cimientos a como estructurar y bajo qué esquema trabajar este modelo de clasificación y agrupación de jugadores de la NBA, por ende también se definirá las diferentes técnicas que los componen esta metodología, además se explicaran los términos y conceptos requeridos para poder comprender el fenómeno bajo el que este está siendo implementado y también se hará mención a las diferentes herramientas tecnológicas para el desarrollo del mismo.

## Antecedentes de la investigación

1. **Análisis discriminante de las estadísticas de juego entre bases, aleros y pivots en baloncesto masculino**

En este artículo sus dos autores (Miguel Ángel Gómez y Alberto Lorenzo Calvo) en el año 2007 para el Instituto Nacional de Educación Física de Catalunya buscaron analizar las diferencias existentes en las estadísticas de juego entre las posiciones de los jugadores en la Euroleague, fue basado en las estadísticas oficiales de la temporada 2004-2005, este trabajo fue de gran utilidad para poder tener un indicio de como una técnica multivariante como lo es el Análisis Discriminante puede ser aplicado para clasificar jugadores de baloncesto de ligas profesionales.

1. **Agrupación y equilibrio competitivo en el baloncesto profesional NBA y ACB**

A su vez en el artículo presentado por múltiples autores para el año 2016 en el mismo instituto se aplicó el método de agrupación (Metodo de Cluster) para observar cómo se podrían separar los equipos dado su nivel de competición.

## Bases Teóricas

Las bases teóricas de esta investigación proporcionarán el fundamento conceptual y metodológico necesario para abordar el análisis del rendimiento de los jugadores de la NBA. Se explorarán teorías y conceptos clave en el ámbito del análisis multivariante y estadística deportiva, con un enfoque particular en los métodos utilizados para evaluar el rendimiento y clasificar a los jugadores. Esto incluirá una revisión de la literatura sobre el análisis de componentes principales (PCA), el análisis de clusters, el análisis discriminante y el análisis multivariante de la varianza (MANOVA). Además, se describirá el contexto específico de la NBA y las posiciones de los jugadores, para entender cómo estas variables influyen en las métricas de rendimiento.

* **National Basketball Association (NBA):**

La National Basketball Association, conocida como NBA, es la principal liga profesional de baloncesto en los Estados Unidos y una de las ligas deportivas más prestigiosas y competitivas del mundo. Fundada en 1946, la NBA está compuesta por 30 equipos, divididos en dos conferencias (Este y Oeste) y varias divisiones. La liga es conocida por su alto nivel de competencia, talento internacional y su capacidad para atraer a una audiencia global. La NBA no solo tiene un impacto significativo en el deporte del baloncesto, sino también en la cultura popular y la economía deportiva global.

* **Posiciones de los Jugadores de Baloncesto:**

En el baloncesto, los jugadores se distribuyen en diferentes posiciones según sus habilidades, roles y responsabilidades en la cancha. Las principales posiciones son:

1. Guard (Base o Escolta):
   * Base (Point Guard, PG): Es el principal manejador del balón y organizador del juego. Su función principal es dirigir las jugadas ofensivas, distribuir el balón a sus compañeros y crear oportunidades de anotación. También se espera que tenga una buena visión de juego y habilidades de dribbling.
   * Escolta (Shooting Guard, SG): Es el encargado de anotar puntos, especialmente desde largas distancias. Los escoltas suelen ser buenos tiradores y tienen la capacidad de crear sus propios tiros. También ayudan en la defensa perimetral.
2. Forward (Alero o Ala-Pívot):
   * Alero (Small Forward, SF): Es un jugador versátil que puede jugar tanto en posiciones exteriores como interiores. Tiene habilidades de anotación, defensa y rebote, y a menudo es responsable de marcar a los jugadores más atléticos del equipo contrario.
   * Ala-Pívot (Power Forward, PF): Juega cerca del aro y se enfoca en anotar desde el poste bajo, recoger rebotes y defender a jugadores interiores. Los ala-pívots necesitan ser fuertes físicamente y tener habilidades de tiro de media distancia.
3. Center (Pívot):
   * Pívot (Center, C): Es el jugador más alto del equipo y juega cerca del aro. Sus responsabilidades incluyen anotar puntos en la pintura, recoger rebotes, bloquear tiros y proteger la canasta. Los pívots deben ser fuertes, buenos en el juego aéreo y tener una presencia defensiva dominante.

* **Métricas de Uso y Rendimiento**
  + El PIE es una métrica avanzada desarrollada por la NBA para medir el impacto global de un jugador en el partido. Se calcula utilizando una fórmula que incluye varias estadísticas de juego. Generalmente se expresa como:
  + El Porcentaje de Uso (USG%) mide el porcentaje de jugadas de un equipo en las que un jugador participa, ya sea lanzando, asistiendo, o perdiendo el balón, mientras está en la cancha. La formula es:
* **Enfoque Multivariante**

El presente trabajo se sustenta en los métodos estadísticos pertenecientes a la rama del análisis multivariante, una disciplina que permite estudiar y comprender las relaciones entre múltiples variables simultáneamente. A continuación, se presentan breves definiciones de los métodos principales que se utilizarán en esta investigación:

* Análisis de Componentes Principales (PCA): El Análisis de Componentes Principales (PCA) es una técnica utilizada para reducir la dimensionalidad de un conjunto de datos. Esto se logra transformando las variables originales en un nuevo conjunto de variables no correlacionadas, conocidas como componentes principales, que capturan la mayor parte de la variabilidad presente en los datos originales. El objetivo del PCA es simplificar los datos sin perder información significativa, facilitando su interpretación y análisis.
* Análisis Multivariante de la Varianza (MANOVA): El Análisis Multivariante de la Varianza (MANOVA) es una extensión del análisis de varianza (ANOVA) que permite evaluar diferencias entre grupos en múltiples variables dependientes simultáneamente. MANOVA analiza si las medias de varias variables dependientes son diferentes entre los niveles de una o más variables independientes categóricas.
* Análisis de Clusters: El Análisis de Clusters es un conjunto de técnicas que agrupan un conjunto de objetos de manera que los objetos dentro de un mismo grupo (o cluster) sean más similares entre sí que con los objetos de otros grupos. Este método es útil para identificar patrones o estructuras en los datos sin necesidad de información previa sobre los grupos.
* Modelo de Mezcla Gaussiana: Es una técnica de modelado estadístico que asume que los datos son una combinación de varias distribuciones gaussianas (normales) con parámetros desconocidos. GMM es una generalización del análisis de clusters, que permite modelar la distribución de datos complejos mediante una combinación ponderada de múltiples gaussianas.
* Análisis Discriminante: El Análisis Discriminante es una técnica utilizada para clasificar observaciones en categorías predefinidas basándose en un conjunto de variables predictoras. Este método busca encontrar una combinación lineal de las variables que mejor separe las diferentes categorías.
* **Herramientas y Tecnologías Utilizadas**

Para llevar a cabo este análisis multivariante del rendimiento de los jugadores de la NBA, se emplearán diversas herramientas y tecnologías de programación y análisis de datos. A continuación, se presentan las herramientas principales y sus roles específicos en esta investigación:

### Lenguajes de Programación (R)

R es un entorno de programación y un lenguaje diseñado específicamente para estadística y análisis de datos. Su robusta comunidad y la amplia gama de paquetes especializados lo convierten en una elección poderosa para la creación y evaluación de modelos estadísticos como PCA, MANOVA, análisis de clusters y análisis discriminante. La flexibilidad de R permite manejar grandes volúmenes de datos y realizar visualizaciones avanzadas con bibliotecas como ggplot2 y tidyverse. Además, su integración con herramientas como RStudio facilita el desarrollo de proyectos analíticos completos, asegurando un flujo de trabajo eficiente y reproducible en el ámbito de la ciencia de datos y la estadística aplicada.

### Entornos de Desarrollo Integrado (IDE-Rstudio)

RStudio es un entorno de desarrollo integrado (IDE) para R, que también soporta la integración con otros lenguajes de programación como Python. se utilizará en combinación con la librería RMarkdown para documentar y presentar el análisis de manera reproducible y comprensible. El entorno facilita la escritura de código, ejecución y depuración, así como la generación de informes interactivos.

### Librerías y Herramientas de Integración

* Rmarkdown: Es una herramienta dentro de RStudio que permite crear documentos dinámicos, combinando código, texto y resultados de análisis en un solo archivo. Se utilizará para documentar el proceso de análisis y generar informes que incluyan tanto el código como los resultados y las visualizaciones, facilitando la comunicación de los hallazgos de manera clara y estructurada.

## Obtención de Datos

La base de datos utilizada en esta investigación se construyó mediante técnicas de web scraping, empleando el lenguaje de programación Python. El web scraping permitió extraer datos detallados y actualizados de las estadísticas avanzadas de los jugadores desde la página oficial de la NBA. Para facilitar la visualización y el manejo de los datos recopilados a usuarios ajenos a lenguajes de programación o a manejadores de bases de datos, la información extraída se cargó en un archivo Excel.

El uso de Python para el web scraping proporcionó una solución eficiente y automatizada para la recolección de datos, garantizando que la base de datos estuviera completa y actualizada. Las bibliotecas de Python como requests, beautifulsoup4 y pandas fueron fundamentales en este proceso, permitiendo la extracción, procesamiento y almacenamiento de los datos de manera

El código utilizado para realizar el web scraping y construir la base de datos está disponible para su consulta y replicación en el siguiente enlace de GitHub-Gist: [Gist Scraper NBA](https://gist.github.com/jarof13/1bc38656af62cfbca57912ed5e076f5c). Este contiene todas las funciones necesarias y las instrucciones detalladas para ejecutar el proceso de extracción de datos, permitiendo a otros investigadores reproducir y validar los resultados obtenidos en este estudio.

# Desarrollo Análisis y Resultados

En esta sección, se presentan los métodos de análisis aplicados a la base de datos recopilada, junto con los resultados obtenidos. El objetivo es explorar y comprender los patrones y relaciones en las métricas de rendimiento de los jugadores de la NBA, utilizando técnicas estadísticas multivariantes avanzadas. Se iniciara con un análisis descriptivo que proporcionará una visión general de los datos, seguido de la aplicación de un análisis multivariante de la varianza (MANOVA) para evaluar diferencias significativas en las métricas clave. Posteriormente, se empleara el análisis de componentes principales (PCA) para reducir la dimensionalidad de los datos y construir una nueva variable que represente el rendimiento de los jugadores. Esta variable será utilizada en un análisis de clusters para agrupar jugadores con características similares, y finalmente, se aplicará un análisis discriminante para clasificar a los jugadores y evaluar la precisión del modelo predictivo desarrollado.

## Análisis Descriptivo

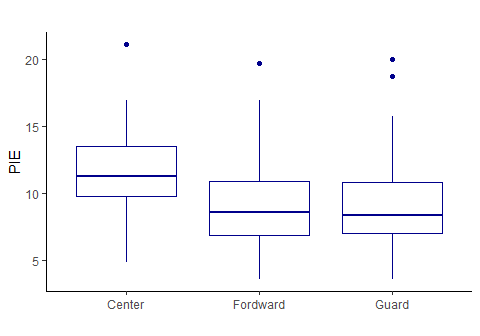
Antes de proceder con el análisis multivariante de la varianza (MANOVA), es fundamental realizar un análisis descriptivo de las métricas clave y las variables categóricas de interés. El análisis descriptivo proporciona una visión general de los datos, permitiendo identificar patrones, tendencias y posibles anomalías que podrían influir en los resultados posteriores. En esta sección, nos enfocamos en las dos métricas principales que se utilizarán en el MANOVA: el Impacto del Jugador Estimado (PIE) y el Porcentaje de Uso (USG%).

### Impacto del Jugador dada la Posición

Se observan variaciones en las métricas estadísticas entre las diferentes posiciones. Los Centros muestran el mayor promedio de impacto global para temporada 2023-2024 indicando potencialmente un impacto más significativo en el juego en comparación con los Aleros y Bases. Sin embargo, estos ultimos se encuentran con una menor dispersión. El rango de valores de PIE varía desde mínimos negativos hasta máximos significativos en todas las posiciones, indicando diversidad en el desempeño de los jugadores.

| Position | Conteo | Media | Desvio | Min | Max | Q1 | Mediana | Q3 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Center | 69 | 11.62 | 2.75 | 4.9 | 21.1 | 9.8 | 11.3 | 13.5 |
| Fordward | 99 | 9.17 | 2.86 | 3.6 | 19.7 | 6.9 | 8.6 | 10.9 |
| Guard | 169 | 9.02 | 2.85 | 3.6 | 20.0 | 7.0 | 8.4 | 10.8 |

Los cuartiles Q1, mediana y Q3 muestran cómo se distribuyen los datos dentro de cada grupo, proporcionando una perspectiva sobre la dispersión y la tendencia central del índice PIE según la posición del jugador en la liga

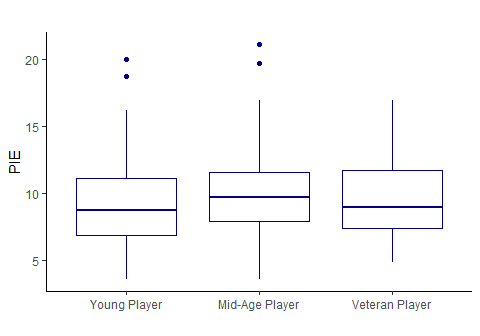


### Impacto del Jugador dado el grupo de Edad

Los jugadores clasificados como “Young Player” (Joven Jugador) muestran una media de PIE ligeramente inferior a la de los “Mid-Age Player” (Jugador de Mediana Edad), pero con una dispersión un poco inferior, como lo indica la desviación estándar. Esto sugiere que los jugadores más jóvenes son los que tienen un impacto en el juego de forma más homogénea. Por otro lado, los “Mid-Age Player” muestran una media de PIE ligeramente superior, lo que podría indicar una mayor consistencia en el rendimiento a medida que los jugadores maduran en sus carreras.

| Age\_Group | Conteo | Media | Desvio | Min | Max | Q1 | Mediana | Q3 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Young Player | 156 | 9.24 | 2.92 | 3.6 | 20.0 | 6.88 | 8.75 | 11.10 |
| Mid-Age Player | 113 | 10.00 | 3.04 | 3.6 | 21.1 | 7.90 | 9.70 | 11.60 |
| Veteran Player | 68 | 9.75 | 3.09 | 4.9 | 16.9 | 7.38 | 9.00 | 11.72 |

Los “Veteran Player” (Jugador Veterano), aunque muestran una media de PIE similar a los jugadores jóvenes, exhiben una dispersión más amplia, como se refleja en la mayor desviación estándar. Esto sugiere que los jugadores veteranos pueden mantener un rendimiento competitivo pero con una variabilidad un poco más significativa en sus contribuciones individuales a lo largo de la temporada.



### Impacto del Jugador dado el Grupo de Edad y la Posición

Las estadísticas descriptivas del índice de Impacto del Jugador Estimado (PIE) condicionadas por la edad y la posición del jugador en la NBA revelan varios patrones interesantes. Los centros jóvenes y veteranos jóvenes son los que poseen un mayor impacto promedio a lo largo de esta temporada, sin embargo este segundo subgrupo es el que posee una mayor dispersión, los armadores jóvenes son los que poseen un menor impacto lo cual parece razonable entendiendo que esa posición requiere de una gran preparación y experiencia de juego la cual van adquiriendo al pasar los años en su carrera, sin embargo es el subgrupo más homogéneo y por ultimo para la temporada en cuestión los aleros jóvenes veteranos fueron el subgrupo de esta posición que más impacto tuvo.

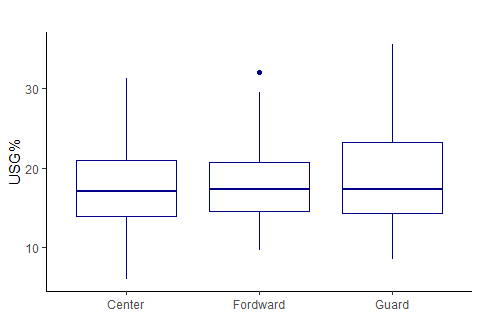
| Age\_Group | Position | Conteo | Media | Desvio | Min | Max | Q1 | Mediana | Q3 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Young Player | Center | 27 | 11.39 | 2.38 | 6.3 | 16.2 | 9.85 | 11.30 | 13.20 |
| Young Player | Fordward | 46 | 8.87 | 2.45 | 4.3 | 14.5 | 6.65 | 8.35 | 10.83 |
| Young Player | Guard | 83 | 8.74 | 3.03 | 3.6 | 20.0 | 6.75 | 8.10 | 10.50 |
| Mid-Age Player | Center | 28 | 12.20 | 2.88 | 7.3 | 21.1 | 9.97 | 11.65 | 13.83 |
| Mid-Age Player | Fordward | 31 | 9.33 | 2.98 | 3.6 | 19.7 | 7.70 | 9.00 | 10.25 |
| Mid-Age Player | Guard | 54 | 9.25 | 2.64 | 4.7 | 15.6 | 7.80 | 8.85 | 10.67 |
| Veteran Player | Center | 14 | 10.91 | 3.12 | 4.9 | 16.8 | 8.58 | 10.65 | 13.28 |
| Veteran Player | Fordward | 22 | 9.60 | 3.49 | 5.4 | 16.9 | 6.98 | 8.50 | 11.67 |
| Veteran Player | Guard | 32 | 9.35 | 2.74 | 5.7 | 15.2 | 7.45 | 8.65 | 11.53 |

### Tasa de Uso del Jugador dada la Posición

Las estadísticas descriptivas del porcentaje de uso (USG%) de los jugadores de la NBA, dadas las posiciones, ofrecen una visión clara sobre cómo se distribuye la carga de trabajo ofensiva entre los diferentes roles en la cancha. Para los Centros, se observa que tienen una media de USG% menor en comparación con los Aleros y los Bases. También es destacable que a pesar de que sea la posición con menor tasa de uso es también la menos homogénea, esto puede dar indicio a que en algunos grupos formados con otras variables los centros poseen un mayor o menor uso.

| Position | Conteo | Media | Desvio | Min | Max | Q1 | Mediana | Q3 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Center | 69 | 17.83 | 5.37 | 6.0 | 31.2 | 13.9 | 17.1 | 21.0 |
| Fordward | 99 | 18.18 | 4.96 | 9.6 | 32.0 | 14.6 | 17.3 | 20.7 |
| Guard | 169 | 18.75 | 5.70 | 8.5 | 35.5 | 14.3 | 17.3 | 23.2 |

Por último, los Bases presentan la media de USG% más alta entre las tres posiciones. Esto refleja su papel central en la ofensiva del equipo, siendo los jugadores que más frecuentemente manejan el balón y realizan jugadas. Ellos tienen una carga de trabajo ofensiva significativamente mayor y pese a ser la posición con mayor cantidad de jugadores es también la que es más homogénea.

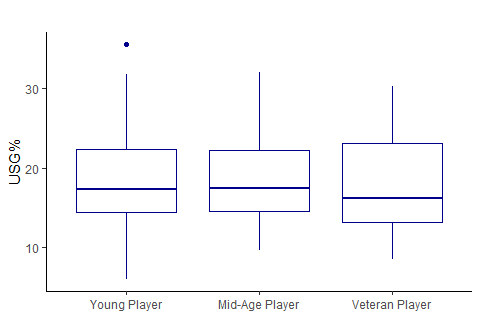


### Tasa de Uso del Jugador dada la Edad

Las estadísticas descriptivas del porcentaje de uso (USG%) de los jugadores de la NBA, categorizados por grupos de edad, proporcionan una comprensión de cómo varía la participación ofensiva según la etapa de la carrera del jugador. Para los “Young Players” (Jugadores Jóvenes), se observa que tienen la media de USG% más alta entre los tres grupos de edad. Esto sugiere que los jugadores jóvenes tienden a tener un papel ofensivo más prominente, posiblemente debido a su energía y frescura física además de ser el grupo de edad con menor dispersión.

| Age\_Group | Conteo | Media | Desvio | Min | Max | Q1 | Mediana | Q3 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Young Player | 156 | 18.46 | 5.30 | 6.0 | 35.5 | 14.47 | 17.3 | 22.30 |
| Mid-Age Player | 113 | 18.63 | 5.44 | 9.6 | 32.0 | 14.60 | 17.4 | 22.20 |
| Veteran Player | 68 | 17.86 | 5.71 | 8.5 | 30.2 | 13.20 | 16.2 | 23.08 |

Para los “Veteran Players” (Jugadores Veteranos), se registra la media de USG% más baja entre los grupos de edad. La mediana y los cuartiles refuerzan esta observación, mostrando que los jugadores veteranos, en general, tienen un menor rol ofensivo.



### Tasa de Uso del Jugador dado el Grupo de Edad y la Posición

Las estadísticas descriptivas del porcentaje de uso (USG%) condicionadas por la edad y la posición del jugador en la NBA revelan varios patrones interesantes. Los guardias jóvenes tienen el mayor porcentaje de uso, reflejando una alta responsabilidad ofensiva desde temprano en sus carreras y la decisión del equipo para poder darle espacio a que se desarrolle. Este grupo también muestra una considerable homogeneidad, indicando una menor variabilidad en su participación ofensiva. Los aleros jóvenes también tienen un porcentaje de uso alto, aunque ligeramente inferior al de los guardias, con una dispersión similar. Los centers jóvenes muestran un menor porcentaje de uso, lo cual puede estar relacionado con las responsabilidades específicas de su posición. A medida que los jugadores envejecen, se observa una disminución en el porcentaje de uso en todas las posiciones. Los centers de mediana edad y veteranos tienen el menor porcentaje de uso, indicando una menor carga ofensiva a medida que progresan en sus carreras. Los forwards y guardias de mediana edad mantienen una participación ofensiva alta, aunque con una leve disminución en comparación con sus contrapartes jóvenes. En general, los guardias, independientemente de la edad, tienden a mantener una alta participación ofensiva, mientras que los centros experimentan la mayor disminución en el porcentaje de uso con la edad.

| Age\_Group | Position | Conteo | Media | Desvio | Min | Max | Q1 | Mediana | Q3 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Young Player | Center | 27 | 17.58 | 5.48 | 6.0 | 31.2 | 14.40 | 17.20 | 19.35 |
| Young Player | Fordward | 46 | 17.97 | 4.44 | 10.4 | 29.2 | 14.22 | 17.55 | 20.28 |
| Young Player | Guard | 83 | 19.02 | 5.67 | 9.7 | 35.5 | 14.70 | 17.30 | 23.25 |
| Mid-Age Player | Center | 28 | 18.88 | 5.44 | 10.1 | 28.8 | 14.65 | 19.05 | 22.65 |
| Mid-Age Player | Fordward | 31 | 18.22 | 5.33 | 11.0 | 32.0 | 14.60 | 17.30 | 20.70 |
| Mid-Age Player | Guard | 54 | 18.73 | 5.59 | 9.6 | 31.1 | 14.60 | 17.35 | 22.15 |
| Veteran Player | Center | 14 | 16.21 | 4.92 | 8.7 | 26.1 | 13.20 | 15.10 | 19.80 |
| Veteran Player | Fordward | 22 | 18.56 | 5.62 | 9.6 | 28.5 | 14.90 | 17.00 | 23.38 |
| Veteran Player | Guard | 32 | 18.11 | 6.10 | 8.5 | 30.2 | 12.78 | 16.15 | 23.40 |

## Modelo Manova

El siguiente paso en este análisis es aplicar el modelo de Análisis Multivariante de la Varianza (MANOVA) para evaluar si existen diferencias significativas en las métricas de rendimiento del Impacto del Jugador Estimado (PIE) y el Porcentaje de Uso (USG%) en función de la posición del jugador y su grupo de edad. El MANOVA es una extensión del ANOVA que permite analizar múltiples variables dependientes simultáneamente, lo que proporciona una visión más integral de cómo los factores categóricos influyen en las métricas de rendimiento. Este análisis permitirá identificar si la posición del jugador y la edad agrupada en categorías tienen un impacto significativo en las métricas de rendimiento mencionadas.

**Hipótesis Nula (H0):**

1. No hay diferencias significativas en PIE y USG% en función de la posición del jugador.
2. No hay diferencias significativas en PIE y USG% en función del grupo de edad del jugador.
3. No hay interacción significativa entre la posición del jugador y su grupo de edad en términos de PIE y USG%.

**Hipótesis Alternativa (H1):**

1. Hay diferencias significativas en PIE y USG% en función de la posición del jugador.
2. Hay diferencias significativas en PIE y USG% en función del grupo de edad del jugador.
3. Hay una interacción significativa entre la posición del jugador y su grupo de edad en términos de PIE y USG%.

|  | Df | Pillai | approx F | num Df | den Df | Pr(>F) |
| --- | --- | --- | --- | --- | --- | --- |
| Position | 2 | 0.28 | 26.69 | 4 | 656 | 0.00 |
| Age\_Group | 2 | 0.02 | 2.03 | 4 | 656 | 0.09 |
| Position:Age\_Group | 4 | 0.01 | 0.50 | 8 | 656 | 0.86 |
| Residuals | 328 | NA | NA | NA | NA | NA |

Para cada una de las hipótesis se tiene:

1. Con un nivel de significación del 5% se puede interpretar que hay diferencias significativas en el Impacto del Jugador Estimado (PIE) y el Porcentaje de Uso (USG%) en función de la posición del jugador. Se puede rechazar la hipótesis nula y concluir que la posición del jugador tiene un impacto significativo en estas métricas de rendimiento.
2. Con un nivel de significación del 5% se puede interpretar que no hay diferencias significativas en el PIE y USG% en función del grupo de edad del jugador. No se encontraron evidencias suficientes para rechazar la hipótesis nula y concluir que la edad del jugador también tiene un impacto significativo en estas métricas.
3. Con un nivel de significación del 5% se puede interpretar que no existe una interacción significativa entre la posición del jugador y su grupo de edad en términos de PIE y USG%. Esto sugiere que el impacto de la posición del jugador en estas métricas no varía según su grupo de edad.

## Análisis de Componentes Principales

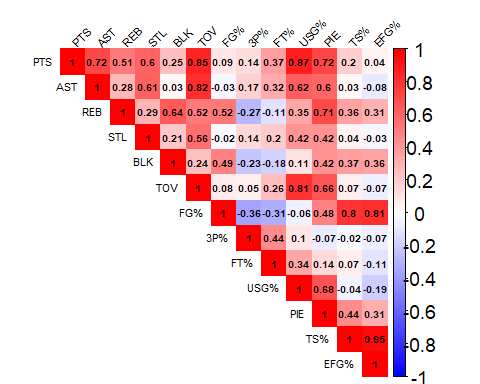
El Análisis de Componentes Principales (PCA) es una técnica estadística utilizada para reducir la dimensionalidad de un conjunto de datos mientras se preserva la mayor cantidad de variabilidad posible. En el contexto del análisis de rendimiento de jugadores de baloncesto de la NBA, PCA permite identificar combinaciones lineales de variables que capturan las características más significativas del rendimiento de los jugadores. Antes de proceder con el PCA, es esencial realizar un estudio de correlación para seleccionar las variables que están más relacionadas con el rendimiento de los jugadores. Este paso asegura que los componentes principales construidos sean representativos de los aspectos más relevantes del juego.

### Selección de Variables para el PCA

Basado en la lógica del deporte y el impacto en el rendimiento de los jugadores, las siguientes variables serán consideradas para el análisis de componentes principales:

* PTS (Puntos): Una métrica directa del rendimiento ofensivo del jugador.
* AST (Asistencias): Refleja la capacidad del jugador para crear oportunidades de anotación para sus compañeros.
* REB (Rebotes): Incluye tanto rebotes ofensivos como defensivos, importante para evaluar la capacidad de control del balón.
* STL (Robos): Indica la habilidad defensiva del jugador.
* BLK (Bloqueos): Otra métrica defensiva importante.
* TOV (Pérdidas de balón): Refleja la eficiencia del jugador en el manejo del balón.
* FG% (Porcentaje de tiros de campo): Evalúa la precisión de los tiros del jugador.
* 3P% (Porcentaje de triples): Evalúa la precisión en tiros de tres puntos.
* FT% (Porcentaje de tiros libres): Mide la eficiencia en los tiros libres.
* USG% (Porcentaje de uso): Indica la frecuencia con la que el jugador está involucrado en las jugadas del equipo.
* PIE (Player Impact Estimate): Una métrica avanzada que estima el impacto del jugador en el juego.
* TS% (True Shooting Percentage): Mide la eficiencia del jugador considerando tiros de campo, triples y tiros libres.
* EFG% (Effective Field Goal Percentage): Ajusta el porcentaje de tiros de campo para dar más peso a los tiros de tres puntos.

Estas variables proporcionan una visión integral tanto del rendimiento ofensivo como defensivo de los jugadores, así como su eficiencia en diferentes aspectos del juego.

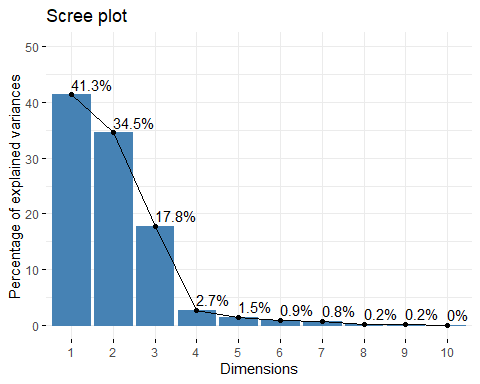


La variable de porcentaje de tiros de tres puntos (3P%) mostró una relación poco significativa con el resto de las variables seleccionadas en el análisis de correlación. Debido a esta baja correlación, se decidió no incluir 3P% en el análisis de componentes principales (PCA). Al priorizar variables que muestran una mayor interrelación, el PCA puede identificar de manera más efectiva los componentes subyacentes que explican la variabilidad en el rendimiento de los jugadores.

## PCA

Al observar el scree plot, se puede notar que los primeros tres componentes principales explican una mayor parte de la varianza en comparación con los componentes posteriores. Haciendo uso del método del porcentaje de varianza explicada que sugiere retener los componentes que, en conjunto, expliquen un porcentaje significativo de la varianza total, generalmente alrededor del 70-80%. En este caso, los primeros tres componentes principales explican una proporción considerable de la varianza total, lo que indica que son suficientes para capturar la mayor parte de la información contenida en los datos originales.

## Warning: In princomp.default(selected\_vars\_pca, scale = TRUE) :  
## extra argument 'scale' will be disregarded



### Interpretación de los Componentes

1. Primer Componente Principal (PC1): El primer componente principal está fuertemente influenciado por las variables FT% (tiros libres) y PTS (puntos). Las cargas de estas variables son positivas, indicando que los jugadores con altas puntuaciones en el primer componente tienden a tener porcentajes altos de tiros libres y anotar más puntos. Variables como USG% (porcentaje de uso) y PIE (impacto del jugador estimado) también tienen cargas positivas, pero menores en comparación con FT% y PTS.
2. Segundo Componente Principal (PC2): El segundo componente principal tiene fuertes cargas positivas en FG% (porcentaje de tiros de campo), TS% (porcentaje de tiros verdaderos) y EFG% (porcentaje de tiros efectivos). Estas variables están asociadas con la eficiencia de los tiros en general. Además, PTS también tiene una carga considerable en este componente, lo que refuerza la idea de que este componente está relacionado con la eficiencia en la anotación. El componente podría ser interpretado como una medida de eficiencia general en el tiro y anotación.
3. Tercer Componente Principal (PC3): El tercer componente principal está influenciado principalmente por PTS, USG%, y REB (rebotes). Esto indica que los jugadores con altas puntuaciones en este componente tienden a ser jugadores que anotan muchos puntos, tienen un alto porcentaje de uso, y recogen una cantidad considerable de rebotes. Las cargas negativas en FT% y TS% sugieren que este componente podría estar capturando un tipo diferente de rendimiento en comparación con los primeros dos componentes, posiblemente relacionado con la contribución general en el juego, incluyendo anotación y rebotes, pero con menos énfasis en la eficiencia de los tiros.

|  | Comp.1 | Comp.2 | Comp.3 |
| --- | --- | --- | --- |
| PTS | 0.36 | 0.40 | 0.47 |
| AST | 0.09 | 0.07 | 0.11 |
| REB | -0.02 | 0.15 | 0.13 |
| STL | 0.01 | 0.01 | 0.02 |
| BLK | -0.01 | 0.02 | 0.01 |
| TOV | 0.04 | 0.04 | 0.06 |
| FG% | -0.36 | 0.56 | -0.05 |
| FT% | 0.75 | 0.16 | -0.59 |
| USG% | 0.32 | 0.22 | 0.47 |
| PIE | 0.06 | 0.23 | 0.17 |
| TS% | -0.12 | 0.44 | -0.26 |
| EFG% | -0.22 | 0.43 | -0.28 |

### Variable de Rendimiento Basado en los Componenetes Principales

La variable “Rendimiento” se ha creado como una combinación ponderada de los tres primeros componentes principales obtenidos del análisis de componentes principales (PCA). Esta técnica se fundamenta en la idea de que los componentes principales capturan la variabilidad de un conjunto de variables originales, condensándola en nuevas dimensiones ortogonales.

En este caso, los componentes principales representan combinaciones lineales de las variables originales que maximizan la varianza explicada en los datos. Al multiplicar cada puntuación de componente principal por su correspondiente autovalor (que representa la varianza explicada por ese componente), y luego sumar estos productos, se obtiene una medida compuesta que refleja la contribución relativa de cada jugador en función de estos componentes fundamentales. Esta nueva variable, “Rendimiento”, está diseñada para capturar de manera sintética las características clave que emergen de los datos originales, permitiendo una evaluación simplificada pero representativa del rendimiento general de los jugadores en la NBA.

| Conteo | Media | Desvio | Min | Max | Q1 | Mediana | Q3 |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 337 | 0 | 64.31 | -223.8 | 162.33 | -43.43 | 0.22 | 40.24 |

## Clustering

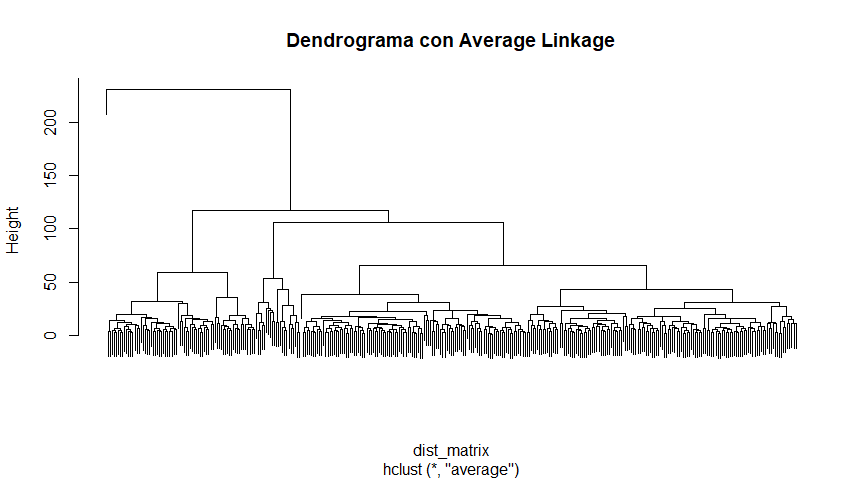
El análisis de clustering es una técnica multivariante utilizada para identificar grupos homogéneos dentro de un conjunto de datos. En el contexto del rendimiento de los jugadores de la NBA, el clustering permite agrupar a los jugadores en segmentos distintos basados en sus características de rendimiento. Este enfoque es especialmente útil para identificar patrones y tendencias que no son evidentes a simple vista y para realizar análisis más profundos sobre subgrupos específicos de jugadores. Para el análisis de clustering, se utilizarán las variables principales obtenidas del análisis de componentes principales (PCA), así como la variable artificial “Rendimiento”, que sintetiza los aspectos más relevantes del rendimiento de los jugadores. Al agrupar a los jugadores de la NBA utilizando estas variables, se espera obtener clusters que representen diferentes perfiles de rendimiento, lo cual puede proporcionar valiosas ideas tanto para el análisis deportivo como para la toma de decisiones estratégicas en equipos y organizaciones.

### Métodos Jerarquicos

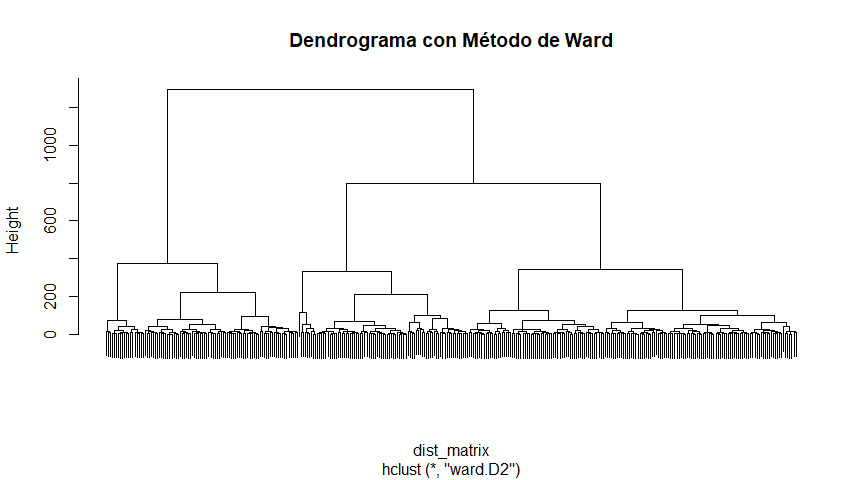
En el análisis de clustering, comenzar con métodos jerárquicos ofrece varias ventajas significativas. Los métodos jerárquicos no requieren que se especifique el número de clusters de antemano, lo que permite una exploración más libre de la estructura de los datos. Este enfoque es especialmente útil en las etapas iniciales del análisis. En particular se usaran:

* Método de Average Linkage: Es una técnica de clustering jerárquico aglomerativo que forma clusters basándose en la media de las distancias entre todos los pares de puntos de dos clusters diferentes.
* Método de Ward: El método de Ward es otro enfoque jerárquico aglomerativo que minimiza la suma de las varianzas dentro de cada cluster.

Observando el dendrograma generado mediante el método de average linkage y realizando un corte en el rango de los pesos entre 50 y 100, se puede determinar de manera preliminar que un número posible de grupos es tres Este análisis inicial se basa en la identificación de ramas largas dentro del dendrograma, las cuales sugieren que los clusters fusionados en esos puntos son significativamente diferentes entre sí. Al trazar una línea horizontal en este intervalo y observar las intersecciones, se obtiene una primera idea del número de clusters que podrían ser utilizados para un análisis más detallado.



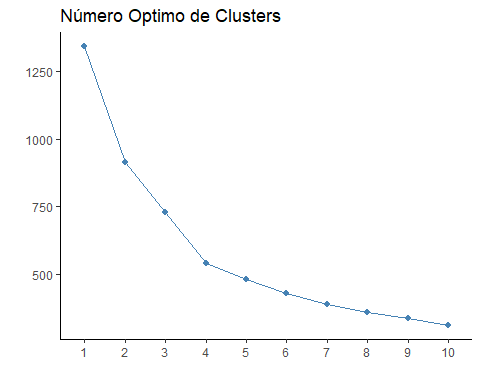
En el dendrograma generado mediante el método de Ward y realizando un corte en el rango de los pesos entre 200 y 300, se puede determinar de igual manera que un número posible de grupos es tres. Este análisis se basa en la tendencia del método de Ward a minimizar la variabilidad dentro de cada cluster, lo que resulta en una estructura de ramas más equilibrada y compacta. Al trazar una línea horizontal en este intervalo y observar las intersecciones, se identifica que la fusión de clusters en estos puntos ofrece una partición natural en tres grupos distintos.



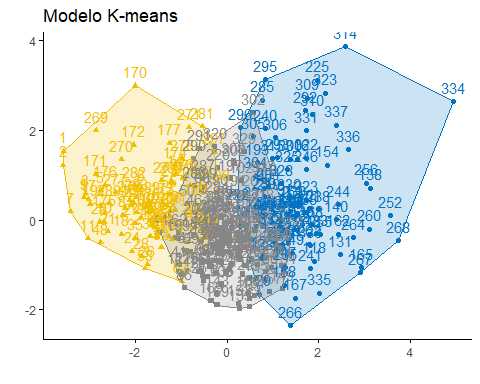
### Métodos No Jerarquicos

Estos métodos son especialmente útiles cuando se tiene una gran cantidad de datos y se requiere una agrupación eficiente y práctica. En el contexto del análisis de jugadores de la NBA, los métodos no jerárquicos permiten identificar grupos de jugadores con perfiles similares en términos de rendimiento y características estadísticas. A diferencia de los métodos jerárquicos, que construyen una jerarquía de clusters de manera ascendente (aglomerativa), los métodos no jerárquicos, como el k-means, se enfocan en particionar el conjunto de datos en un número predefinido de clusters. Este enfoque permite una mayor flexibilidad y escalabilidad, especialmente cuando se trabaja con grandes conjuntos de datos.

Para el análisis de jugadores de la NBA, se seleccionarán las variables más representativas del rendimiento de los jugadores, incluyendo los componentes principales derivados del análisis PCA y la variable artificial de rendimiento. Estas variables cuantitativas permitirán una agrupación más precisa y significativa, destacando patrones y tendencias en el rendimiento de los jugadores. Además, la combinación de métodos jerárquicos y no jerárquicos asegura una evaluación preliminar robusta del número de clusters y una optimización eficiente de los grupos resultantes.



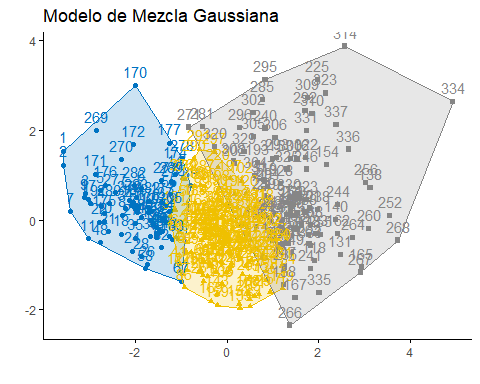
Se utilizó el método del codo para reforzar la decisión de usar de tres a cuatro clusters. Al graficar la WSS para diferentes números de clusters, se observó que la disminución de la WSS se hace menos pronunciada después de cuatro clusters. Este “codo” indica que cuatro clusters son suficientes para capturar la estructura subyacente de los datos sin introducir una complejidad innecesaria.



Como se puede observar el Cluster 1 tiene un rendimiento medio notablemente bajo, con una dispersión considerable, oscilando entre un rango muy amplio desde los. En contraste, el Cluster 2 muestra un rendimiento promedio significativamente más alto, con una dispersión más estrecha en comparación con los otros clusters. Por último, el Cluster 3 tiene un rendimiento medio moderado, con una dispersión considerable y valores que van desde negativos hasta positivos. Estas diferencias sugieren que los clusters representan grupos distintos en términos de rendimiento, siendo el Cluster 2 el más destacado en términos de valores medios elevados y menor variabilidad interna.

| Cluster | Conteo | Media | Desvio | Min | Max | Q1 | Mediana | Q3 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | 97 | -70.97 | 36.37 | -223.80 | -10.03 | -84.64 | -63.43 | -45.39 |
| 2 | 90 | 77.91 | 37.03 | 7.98 | 162.33 | 53.50 | 75.18 | 102.46 |
| 3 | 150 | -0.85 | 26.63 | -64.09 | 66.26 | -19.56 | 1.49 | 17.52 |

Ahora procederemos a realizar la agrupación mediante el Modelo de Mezcla Gaussiana para comparar los grupos obtenidos y determinar cuál método de agrupación ofrece mejores resultados.



Para interpretar los grupos creados mediante el modelo de Mezcla Gaussiana se tiene que el grupo 1 se caracteriza por un alto rendimiento, representado por puntajes promedio significativamente altos y una variabilidad moderada. Los jugadores en este grupo tienden a contribuir consistentemente con puntajes altos en los juegos, mostrando habilidades destacadas en múltiples estadísticas que contribuyen al rendimiento general del equipo, el grupo 2 muestra un rendimiento medio-bajo. Aunque tienen una dispersión moderada en sus puntajes, la media general es negativa, lo que indica que algunos jugadores en este grupo pueden tener contribuciones variables y menos consistentes en comparación con los otros grupos. Por ultimo, el grupo 3 representa el rendimiento más bajo, con puntajes promedio negativos los jugadores en este grupo pueden enfrentar desafíos consistentes en su rendimiento, reflejado en puntajes generalmente más bajos en comparación con los otros grupos.

| GM | Conteo | Media | Desvio | Min | Max | Q1 | Mediana | Q3 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | 73 | 89.48 | 30.86 | 36.40 | 162.33 | 64.56 | 84.23 | 115.92 |
| 2 | 166 | 1.37 | 26.72 | -56.47 | 56.03 | -18.65 | 5.01 | 21.35 |
| 3 | 98 | -68.98 | 39.61 | -223.80 | 37.64 | -84.26 | -63.76 | -45.96 |

Basado en la estructura descriptiva de los grupos obtenidos mediante ambos métodos de clustering, parece ser que el Modelo de Mezcla Gaussiana está proporcionando una agrupación más clara y distintiva. Esto se evidencia en la definición más precisa de los grupos según las características de rendimiento de los jugadores. Mientras que el método K-means mostró tres grupos con variabilidades en el rendimiento, el Modelo de Mezcla Gaussiana ofreció grupos con diferencias más marcadas en términos de rendimiento promedio y variabilidad. Esta observación sugiere que el Modelo de Mezcla Gaussiana podría estar capturando mejor la estructura subyacente de los datos en términos de agrupación basada en el rendimiento de los jugadores.

## Análisis Discriminante

El análisis discriminante es una técnica estadística utilizada para clasificar observaciones en grupos predefinidos basados en diversas características cuantitativas. En el contexto de este estudio, se aplicará el análisis discriminante para evaluar la efectividad de la clasificación de los jugadores de la NBA en los grupos formados por el Modelo de Mezcla Gaussiana. Utilizando el dataframe construido, que incluye variables relevantes como la variable artificial de rendimiento, los componentes principales seleccionados (PC1, PC2 y PC3) y la edad de los jugadores. El objetivo del análisis discriminante es verificar la precisión y validez de las agrupaciones realizadas por el modelo gaussiano, proporcionando una comprensión más profunda de las variables que discriminan entre los diferentes grupos de rendimiento de los jugadores. A través de este enfoque, se espera corroborar la adecuación del Modelo de Mezcla Gaussiana en la segmentación de los jugadores de la NBA y, al mismo tiempo, obtener insights sobre las características distintivas de cada grupo identificado.

En el análisis discriminante lineal (LDA), las medias de cada grupo en el espacio de las variables discriminantes proporcionan información clave sobre las características promedio de cada grupo. A continuación se presenta una interpretación detallada de las medias de cada grupo basado en los resultados obtenidos:

* Componente Principal 1 (PC1):
  + Grupo 1 tiene una media positiva alta en PC1 (11.25), indicando que los jugadores en este grupo tienden a tener altos valores en las variables originales que contribuyen positivamente a este componente.
  + Grupo 2 tiene una media positiva moderada en PC1 (2.67), sugiriendo valores moderadamente altos en las mismas variables.
  + Grupo 3 tiene una media negativa alta en PC1 (-12.90), lo que indica que los jugadores en este grupo tienen bajos valores en las variables que contribuyen a este componente.
* Componente Principal 2 (PC2):
  + Grupo 1 muestra una media positiva alta en PC2 (7.72), indicando una fuerte presencia de las variables originales que influyen positivamente en este componente.
  + Grupo 2 tiene una media negativa en PC2 (-2.49), sugiriendo una influencia negativa de las variables correspondientes.
  + Grupo 3 también muestra una media negativa en PC2 (-1.52), aunque menos pronunciada que en el grupo 2.
* Componente Principal 3 (PC3):
  + Grupo 1 tiene una media positiva en PC3 (6.06), indicando altos valores en las variables que contribuyen a este componente.
  + Grupo 2 tiene una media negativa en PC3 (-3.32), lo que sugiere una influencia negativa de las variables originales en este grupo.
  + Grupo 3 tiene una media ligeramente positiva en PC3 (1.11).
* Rendimiento:
  + Grupo 1 presenta una media de rendimiento muy alta (89.48), indicando que los jugadores en este grupo tienen un rendimiento significativamente superior.
  + Grupo 2 tiene una media de rendimiento casi neutra (1.37), sugiriendo un rendimiento promedio.
  + Grupo 3 muestra una media de rendimiento negativa significativa (-68.98), indicando que los jugadores en este grupo tienen un rendimiento significativamente inferior.
* Edad:
  + Grupo 1 tiene una media de edad de 27.26 años, lo que sugiere que los jugadores en este grupo son ligeramente mayores.
  + Grupo 2 tiene una media de edad de 26.77 años, indicando una edad promedio.
  + Grupo 3 tiene una media de edad de 25.95 años, sugiriendo que los jugadores en este grupo son ligeramente más jóvenes.

| PC1 | PC2 | PC3 | Rendimiento | Age |
| --- | --- | --- | --- | --- |
| 11.25 | 7.72 | 6.06 | 89.48 | 27.26 |
| 2.67 | -2.49 | -3.32 | 1.37 | 26.77 |
| -12.90 | -1.52 | 1.11 | -68.98 | 25.95 |

Los coeficientes de los componentes discriminantes proporcionan una idea de cómo las variables originales contribuyen a la discriminación entre los grupos. En este caso, los coeficientes (scaling) para los primeros dos componentes discriminantes (LD1 y LD2) son los siguientes:

* LD1 (Primera Discriminante Lineal):
  + La variable PC1 tiene el mayor coeficiente (0.07) en LD1, indicando que PC1 es el componente más significativo para la discriminación en esta dirección.
  + PC3 también contribuye de manera importante (0.06), seguido por Rendimiento (0.02)
* LD2 (Segunda Discriminante Lineal):
  + PC3 tiene un coeficiente negativo significativo (-0.13), lo que indica que es un factor importante en la discriminación en la dirección opuesta a LD1.
  + PC1 tiene un coeficiente positivo pequeño (0.07), lo que sugiere que también contribuye a LD2, pero en menor medida.

Este análisis permite concluir que PC1 y PC3 son las variables clave para discriminar entre los grupos en esta muestra. PC2 y Age también juegan un papel, aunque menos importante. Rendimiento es relevante solo en la primera discriminante lineal (LD1).

|  | LD1 | LD2 |
| --- | --- | --- |
| PC1 | 0.07 | 0.07 |
| PC2 | 0.01 | -0.04 |
| PC3 | 0.06 | -0.13 |
| Rendimiento | 0.02 | 0.00 |
| Age | 0.01 | -0.03 |

Los valores singulares (SVD) en el contexto del análisis discriminante lineal (LDA) representan la relación de varianza explicada por cada una de las funciones discriminantes. Los valores se interpretan como:

* LD1: Este es el valor singular correspondiente a la primera función discriminante, representa la mayor parte de la varianza explicada por el modelo y es la función discriminante más importante para separar los grupos.
* LD2: Este es el valor singular correspondiente a la segunda función discriminante y explica una parte significativa pero menor de la varianza en comparación con LD1.

| Function | Singular\_Value |
| --- | --- |
| LD1 | 24.82 |
| LD2 | 9.21 |

Estos resultados sugieren que LD1 es la principal dimensión de discriminación entre los grupos, mientras que LD2 añade una dimensión adicional pero menos dominante.Si los valores singulares son significativamente distintos de cero, indican que las funciones discriminantes están explicando una porción significativa de la varianza en los datos. Valores muy cercanos a cero podrían sugerir una pobre separación entre grupos.

### Matriz de Confusión

La matriz de confusión muestra cómo el modelo de análisis discriminante clasifica las observaciones en cada grupo. Los resultados se presentan a continuación:

| Actual/Predicción | **Grupo 1** | **Grupo 2** | **Grupo 3** |
| --- | --- | --- | --- |
| **Grupo 1** | 0 | 72 | 0 |
| **Grupo 2** | 18 | 18 | 144 |
| **Grupo 3** | 79 | 0 | 6 |

* **Grupo 1**: Se clasificaron correctamente 0 observaciones como pertenecientes al grupo 1, 72 observaciones fueron clasificadas incorrectamente como grupo 2, y 0 como grupo 3.
* **Grupo 2**: Se clasificaron correctamente 18 observaciones como pertenecientes al grupo 2, 18 observaciones fueron clasificadas incorrectamente como grupo 1, y 144 como grupo 3.
* **Grupo 3**: Se clasificaron correctamente 6 observaciones como pertenecientes al grupo 3, 79 observaciones fueron clasificadas incorrectamente como grupo 1, y 0 como grupo 2.

Basado en los resultados de la matriz de confusión obtenida, el modelo mostró dificultades significativas en la clasificación precisa de todas las observaciones en sus respectivos grupos. Los errores podrían indicar que las variables utilizadas en el modelo no capturan completamente las diferencias entre los grupos, o que existen interacciones complejas entre las variables que el modelo no pudo captar adecuadamente. Basado en estos resultados, se puede concluir que el modelo de análisis discriminante aplicado en su forma actual no es adecuado para clasificar con precisión los jugadores en los grupos definidos por las variables utilizadas.

## Conclusiones

Se observaron diferencias significativas en el Impacto del Jugador Estimado (PIE) y el Porcentaje de Uso (USG%) según la posición del jugador, indicando un impacto diferencial en estas métricas de rendimiento entre las posiciones. En contraste, no se encontraron diferencias significativas en el PIE y USG% según el grupo de edad del jugador, sugiriendo que la edad no influye de manera relevante en estas métricas. Además, no se evidenció una interacción significativa entre la posición del jugador y su grupo de edad en términos de PIE y USG%, lo que sugiere que el efecto de la posición del jugador en estas métricas no varía con la edad.

Se aplicó un Análisis de Componentes Principales (PCA) a un conjunto de métricas clave del rendimiento de jugadores de baloncesto. Se redujo la dimensionalidad a 3 componentes principales que explican más del 80% de la variabilidad de las variables originales. El PC1 está dominado por FT% (porcentaje de tiros libres) y PTS (puntos), destacando la importancia de la precisión en tiros libres y la capacidad de anotación. El PC2 se caracteriza por FG% (porcentaje de tiros de campo), TS% (porcentaje de tiros verdaderos) y EFG% (porcentaje de tiros efectivos), reflejando la eficiencia general en el tiro. El PC3 se centra en PTS, USG% (porcentaje de uso) y REB (rebotes), indicando un perfil de jugador que anota, participa frecuentemente en jugadas y obtiene rebotes. Posteriormente, se creó una variable de rendimiento basada en estos componentes principales. Es importante destacar que ninguna métrica defensiva tuvo especial relevancia al momento de crear los componentes, lo que puede afectar a los jugadores que se destacan o aportán más en el ámbito defensivo, señalando una debilidad en el enfoque al priorizar el aspecto ofensivo del juego.

Mediante el modelo de mezcla gaussiana se formaron tres grupos distintos de jugadores. Este método resultó en grupos coherentes, destacándose especialmente el primer grupo, donde se agruparon jugadores con un alto rendimiento ofensivo, tal como fue captado por los componentes principales y la variable artificial de rendimiento. Estos hallazgos sugieren que el modelo de mezcla gaussiana fue efectivo para identificar perfiles distintivos de jugadores basados en su desempeño en el campo, proporcionando una agrupación que podría ser útil para estrategias de evaluación y desarrollo de talento en el baloncesto.

El análisis discriminante realizado utilizando los grupos definidos por el modelo de mezcla gaussiana mostró dificultades significativas en la clasificación precisa de todas las observaciones en sus respectivos grupos. Los resultados obtenidos de la matriz de confusión revelaron que el modelo no fue eficiente en la clasificación de los jugadores. Los errores observados sugieren posibles limitaciones en las variables utilizadas, que podrían no capturar completamente las diferencias entre los grupos formados, o indicar la presencia de interacciones complejas entre las variables que el modelo no pudo manejar adecuadamente. En conclusión, basándose en estos hallazgos, se puede afirmar que el modelo de análisis discriminante aplicado en su configuración actual no es adecuado para clasificar con precisión los jugadores en los grupos definidos por las variables empleadas.

## Bibliografía

1. Aluja, T., Morineau, A. (1999). *Aprender de los datos: el análisis de componentes principales, una aproximación desde el data mining*. EUB Barcelona.
2. Anderson T. (2003). *An Introduction to Multivariate Statistical Analysis*. Willey Series in Probability and Statistics. New Jersey.
3. Canavos, George (2003) *Probabilidad y estadística. Aplicaciones y métodos*. McGraw-Hill, México.
4. Cuadras C. M. (2019). *Nuevos Métodos de Análisis Multivariante*. CMC Editions Barcelona.
5. García Pérez, A. (2005). *Métodos avanzados de estadística aplicada. Técnicas avanzadas*. UNED. Madrid.
6. Johnson, D. E. (2000). *Métodos multivariados aplicados al Análisis de Datos*. International Thomson Editores. México.
7. Johnson A., Wichern D.W. (2007). *Applied Multivariate Statistical Analysis*. Pearson.
8. Peña, D. (2002). *Análisis de datos multivariantes*. Mc Graw Hill Interamericana. Madrid. España.
9. Pérez, C. (2008). *Técnicas de Análisis Multivariante de Datos. Aplicaciones con SPSS*. Pearson. Madrid.