



Universidad de Oviedo

FACULTAD DE CIENCIAS

ANÁLISIS DE DATOS

ANÁLISIS Y PREDICCIÓN DE LOS SALARIOS EN LA NBA EN BASE AL RENDIMIENTO ESTADÍSTICO DE LOS JUGADORES TEMPORADA 2022/23

GRUPO 4

DE SOTO MANILLA, DAVID

LLORIÁN GONZÁLEZ, PABLO

MARTÍNEZ JÍMENEZ, YAIZA

MENÉNDEZ FERNÁNDEZ, CRISTINA

PÉREZ NAVARRO, DAVID

5 de mayo de 2025

Índice

1	Introducción	3
2	Resultados	4
2.1	Preprocesamiento y transformación de variables	4
2.2	Análisis de componentes principales (PCA)	5
2.3	Relación entre eficiencia y salario	9
2.4	Análisis de agrupamiento: Clustering defensivo	11
2.5	Red neuronal para la predicción del salario	14
2.6	Relación entre la posición en el campo y el salario	18
2.7	Análisis del salario por minuto	19
2.8	Relación entre el perfil de tiro y el salario	20
3	Conclusiones	25
4	Reparto de tareas	26
5	Bibliografía	27
A	Código	28

1. Introducción

En este trabajo se analiza la relación entre el rendimiento estadístico de los jugadores de la NBA y sus salarios, con el objetivo de comprender los factores que influyen en la retribución económica dentro de este contexto profesional. Para ello, se ha utilizado una base de datos que recoge información detallada sobre estadísticas individuales de los jugadores, incluyendo puntos, asistencias, rebotes, tapones, robos y datos contractuales como el salario anual, durante la temporada 2022/23.

El propósito principal del estudio es identificar patrones y relaciones entre variables de rendimiento y salario, así como construir modelos que permitan predecir la retribución de un jugador a partir de sus métricas de juego. Además, se exploran diferencias entre perfiles de jugador (según su posición en el campo o estilo defensivo) y se evalúan posibles casos de sobrevaloración o infravaloración salarial.

Desde el punto de vista metodológico, se ha llevado a cabo un preprocesamiento de los datos, incluyendo limpieza, transformación de variables y creación de nuevas métricas como la eficiencia. Posteriormente, se aplicaron técnicas estadísticas multivariantes como el análisis de componentes principales (PCA), pruebas de correlación, análisis de agrupamiento (clustering), análisis de correspondencias y modelos predictivos mediante redes neuronales y árboles de regresión. Las herramientas utilizadas han sido R y Python (a través de Google Colab).

Este enfoque combinado ha permitido obtener una visión completa y estructurada del estudio que nos planteábamos. En las secciones siguientes se detallan los resultados obtenidos en cada una de las fases del estudio.

Para introducir la parte troncal del trabajo realizado, exponemos a continuación las principales variables consideradas en el análisis:

- **Name:** Nombre y apellido del jugador.
- **Salario:** Salario.
- **POS:** Posición en el campo.
- **Age:** Edad.
- **GP:** Número de partidos jugados en la temporada.
- **Min:** Minutos jugados.
- **PTS:** Puntos totales.
- **FGM:** Tiros de campo encestandos.
- **FGA:** Tiros de campo intentados.
- **FG:** Porcentaje de acierto en tiros de campo.
- **X3PM:** Triples encestandos.

- **X3PA:** Triples intentados.
- **X3P %:** Porcentaje de acierto en triples.
- **FT:** Tiros libres encestados.
- **FTA:** Intentos de tiros libres.
- **FT %:** Porcentaje de acierto en tiros libres.
- **AST:** Asistencias.
- **TOV:** Pérdidas.
- **STL:** Robos.
- **BLK:** Tapones.
- **PF:** Faltas personales.
- **DD2:** Dobles dobles.
- **TD3:** Triples dobles.
- **Grupo_Salario:** Salarios agrupados en rangos.
- **Eficiencia:** Medida de rendimiento.
- **PTS_por_partido:** Puntos por partido jugado.
- **AST_por_partido:** Asistencias por partido jugado.
- **REB_por_partido:** Rebotes por partido jugado.
- **BLK_por_partido:** Tapones por partido jugado.
- **STL_por_partido:** Robos por partido jugado.

2. Resultados

2.1. Preprocesamiento y transformación de variables

En primer lugar, llevamos a cabo un proceso de limpieza de la base de datos original. Identificamos cuatro jugadores cuya posición figuraba como N/A y que solamente habían disputado entre uno y cuatro partidos durante toda la temporada. Dado que su aportación estadística era irrelevante, decidimos eliminarlos del conjunto de datos para evitar posibles sesgos.

A continuación, transformamos la variable *Salario*, originalmente expresada en dólares, a millones de dólares, para así facilitar su interpretación. A partir de esta variable, construimos una nueva

variable categórica, denominada *Grupo_Salario*, que clasifica a los jugadores en intervalos de 10 millones: 0-10M, 10-20M, 20-30M, 30-40M y 40-50M.

Para entender mejor el rendimiento de cada jugador, calculamos sus estadísticas por partido. Es decir, dividimos sus totales de puntos, asistencias, rebotes, tapones y robos entre el número de partidos que jugaron. Así, obtuvimos las nuevas variables:

- Puntos por partido (*PTS_por_partido*)
- Asistencias por partido (*AST_por_partido*)
- Rebotes por partido (*REB_por_partido*)
- Tapones por partido (*BLK_por_partido*)
- Robos por partido (*STL_por_partido*)

Por último, incorporamos una métrica avanzada denominada *eficiencia*, calculada según la siguiente fórmula empleada habitualmente en el análisis estadístico de la NBA:

$$\text{Eficiencia} = \frac{\text{PTS} + \text{REB} + \text{AST} + \text{STL} + \text{BLK} - (\text{FGA} - \text{FGM}) - (\text{FTA} - \text{FTM}) - \text{TOV}}{\text{GP}}$$

donde cada abreviatura representa estadísticas individuales del jugador, y *GP* es el número de partidos disputados.

Gracias a este preprocesamiento, contamos con una base de datos más limpia, coherente y enriquecida, lo cual nos facilita tanto el análisis exploratorio como la implementación de modelos predictivos en las siguientes fases del trabajo.

2.2. Análisis de componentes principales (PCA)

Para reducir la dimensión de los datos y detectar patrones en el rendimiento de los jugadores, realizamos un análisis de componentes principales (PCA) utilizando únicamente las variables numéricas (excluyendo el salario). El PCA se llevó a cabo con un centrado y escalado de las variables.

Componente	Varianza explicada (%)	Varianza acumulada (%)
PC1	54.5	54.5
PC2	11.1	65.6
PC3	6.9	72.6

Tabla 1: Porcentaje de varianza explicada por las tres primeras componentes principales

Como se observa en la Tabla 1, las tres primeras componentes explican más del 72 % de la variabilidad total. En la Figura 1 se visualiza cómo decae la varianza explicada a medida que se incorporan

más componentes.

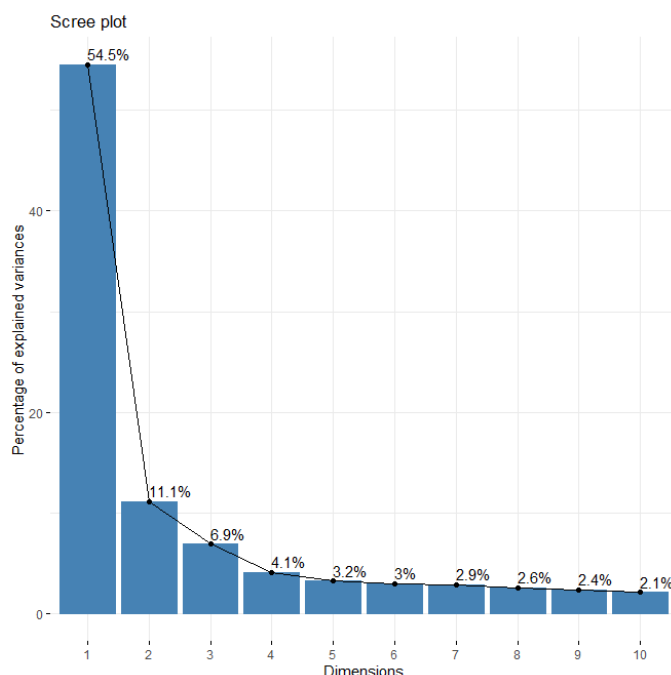


Figura 1: Porcentaje de varianza explicada por cada componente

A partir de la interpretación de las cargas de las variables en cada componente, concluimos lo siguiente:

- **PC1 (54.5 % de varianza):** Representa el rendimiento total del jugador. Incluye minutos jugados, puntos, asistencias, rebotes y eficiencia. Distingue claramente a los jugadores titulares con mayor impacto que el resto.
- **PC2 (11.1 % de varianza):** Se relaciona con el estilo de juego. Distingue a los jugadores interiores (más rebotes y tapones y mejor porcentaje de acierto en tiros de campo) de los jugadores exteriores (más triples intentados y encestados).
- **PC3 (6.9 % de varianza):** Refleja la versatilidad del jugador, capturando la capacidad de conseguir dobles dobles o triples dobles, así como asistencias y robos. Representa jugadores con un impacto equilibrado en el juego.

La Figura 2 nos muestra la representación de las variables sobre los dos primeros ejes del PCA. La dirección de las flechas indica la magnitud de la contribución de cada variable a las componentes principales. Cuanto más larga y alineada con un eje esté una flecha, mayor es su relación con esa componente.

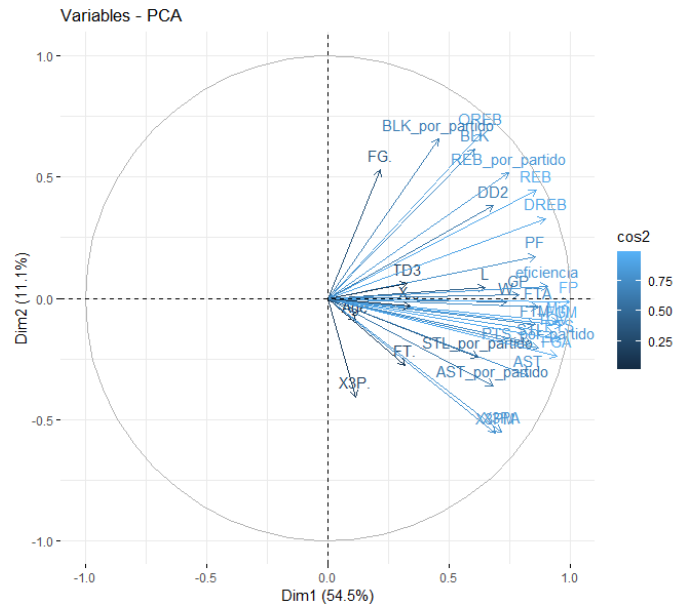


Figura 2: Proyección de las variables en el plano de las dos primeras componentes principales.

También analizamos la relación entre estas tres componentes y el salario. Obtuvimos las siguientes correlaciones:

- Correlación salario–PC1: **0.67**
- Correlación salario–PC2: **-0.11**
- Correlación salario–PC3: **0.31**

Estos resultados indican que el salario está principalmente asociado al rendimiento total (PC1), seguido en menor medida por la versatilidad (PC3). El estilo de juego (PC2) apenas influye.

En la Figura 3, mostramos a los jugadores proyectados sobre las dos primeras componentes principales, coloreados según su salario. Se observa claramente que los jugadores con mayores valores en la primera componente principal tienden a tener salarios más altos.

Este análisis muestra que el rendimiento total del jugador influye mucho en el salario, por lo que las componentes principales pueden ser útiles para ayudar a predecirlo en futuros modelos.

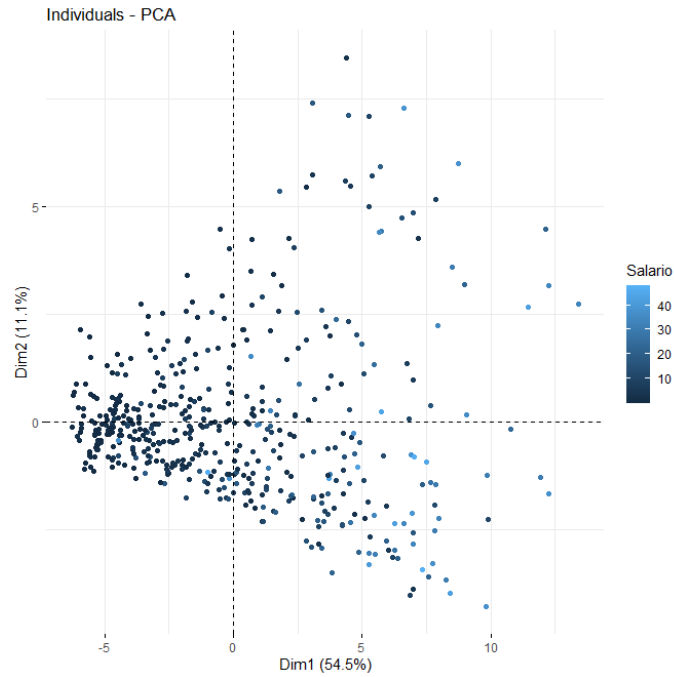


Figura 3: Representación de los jugadores en el espacio de las componentes principales, coloreados en función del salario.

Para completar el análisis visual, representamos a los jugadores en un espacio tridimensional formado por las tres primeras componentes principales. Utilizamos la librería `plotly` de R, que permite realizar gráficos interactivos en tres dimensiones.

En la Figura 4 y la Figura 5, mostramos dos proyecciones distintas de este espacio 3D. Se aprecia cómo los jugadores con mayores salarios (colores más claros) tienden a agruparse en regiones con altos valores en PC1 y PC3, lo que refuerza la idea de que estas componentes recogen información relevante para explicar el salario.

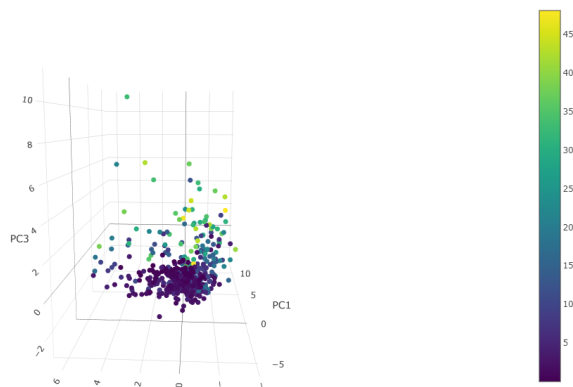


Figura 4: Visualización 3D de los jugadores según las tres primeras componentes principales (vista 1). El color representa el salario.

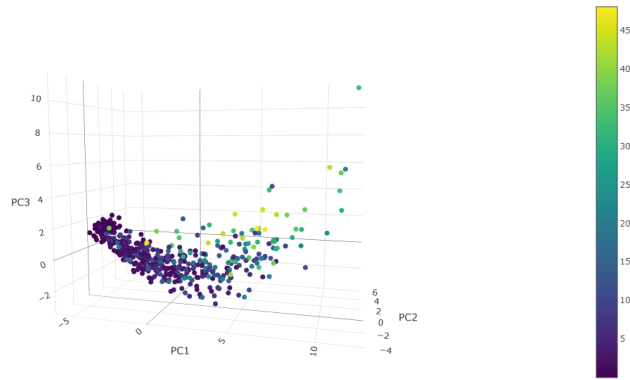


Figura 5: Visualización 3D de los jugadores según las tres primeras componentes principales (vista 2). El color representa el salario.

2.3. Relación entre eficiencia y salario

Comenzamos el análisis de la variable *Salario*, representada en millones de dólares. Para facilitar su interpretación, utilizaremos la variable *Grupo Salario* que comentamos antes. En la Figura 6 observamos que la mayoría de los jugadores pertenece al grupo de menor salario (0–10M), mientras que los contratos más elevados son mucho menos frecuentes.

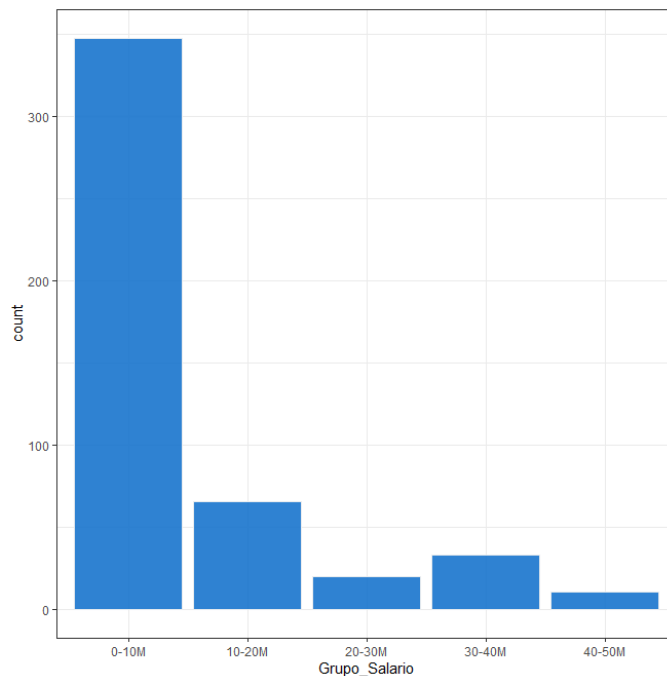


Figura 6: Distribución del número de jugadores por grupo salarial

Posteriormente, analizamos la relación entre el salario y la eficiencia estadística de los jugadores. Como se muestra en la Figura 7, existe una clara tendencia creciente: los jugadores con mayor

eficiencia tienden a tener salarios más altos. Además, destacamos con puntos rojos a los dos jugadores más relevantes: **Nikola Jokić**, el más eficiente, y **Stephen Curry**, el mejor pagado.

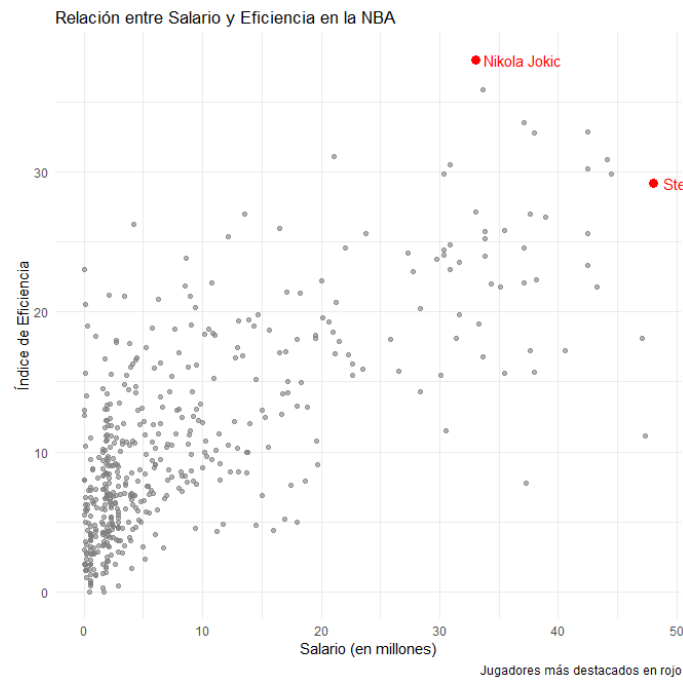


Figura 7: Relación entre salario y eficiencia. Jugadores destacados en rojo

Para visualizar esta relación de forma agrupada, elaboramos un diagrama de cajas que compara la eficiencia entre los distintos grupos salariales. Tal como se aprecia en la Figura 8, existe una relación positiva: los jugadores de mayor salario suelen presentar índices de eficiencia más elevados.

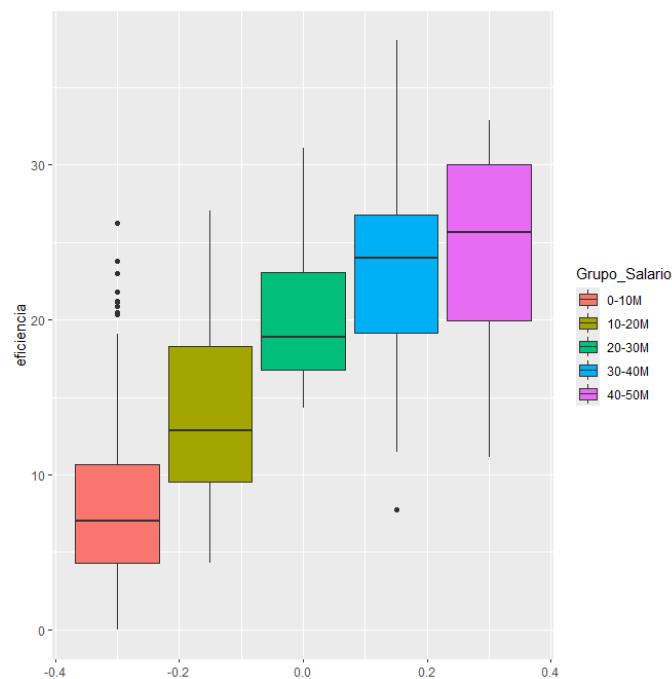


Figura 8: Distribución de la eficiencia según los grupos salariales

Finalmente, realizamos un test de correlación de Pearson entre el salario y la eficiencia, obteniendo un coeficiente de 0.73, con un p-valor menor que 2.2×10^{-16} , lo que indica una relación estadísticamente significativa. Rechazamos así la hipótesis nula de independencia entre ambas variables.

Este análisis muestra que existe una fuerte asociación entre el rendimiento individual (medido por la eficiencia) y el salario de los jugadores, como cabía esperar.

2.4. Análisis de agrupamiento: Clustering defensivo

Aplicamos el algoritmo no jerárquico de las k-medias para agrupar a los jugadores según sus estadísticas defensivas, concretamente robos (*STL_por_partido*) y tapones por partido (*BLK_por_partido*). Establecimos tres grupos, que se interpretan como tipos de defensores: malos defensores, buenos taponadores y buenos ladrones.

La Figura 9 muestra el resultado del agrupamiento en el plano definido por las dos variables, con los tres clústeres bien diferenciados.

Al analizar los grupos formados, observamos que el grupo 1 (rojo) destaca por su media más alta en robos por partido, aunque con pocos tapones, lo que nos lleva a interpretarlo como el grupo de ladrones. El grupo 2 (azul) presenta bajos promedios en ambas métricas, por lo que se considera el de malos defensores. En cambio, el grupo 3 (verde) se caracteriza por un alto promedio de tapones y pocos robos, siendo clasificado como el grupo de taponadores.

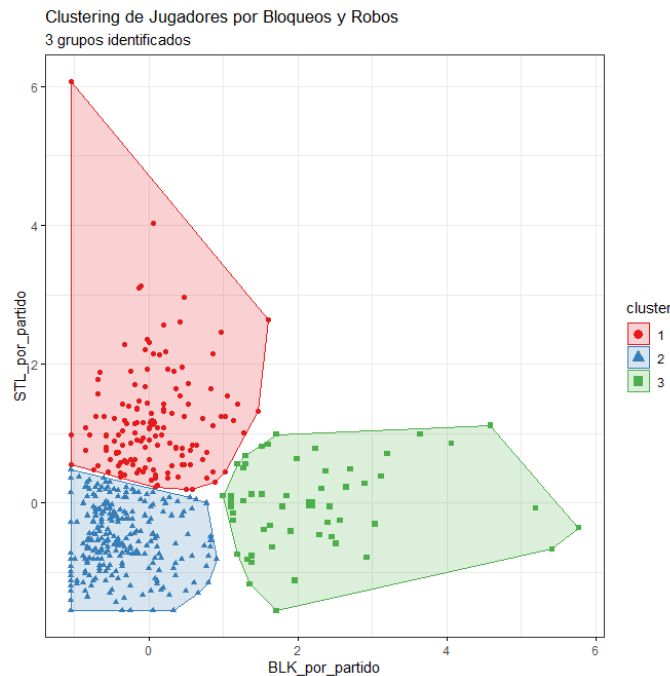


Figura 9: Clustering de jugadores en función de robos y tapones por partido

Estas observaciones se confirman visualmente en las Figuras 10 y 11, que muestran los diagramas

de caja para tapones y robos por partido, respectivamente, en cada grupo.

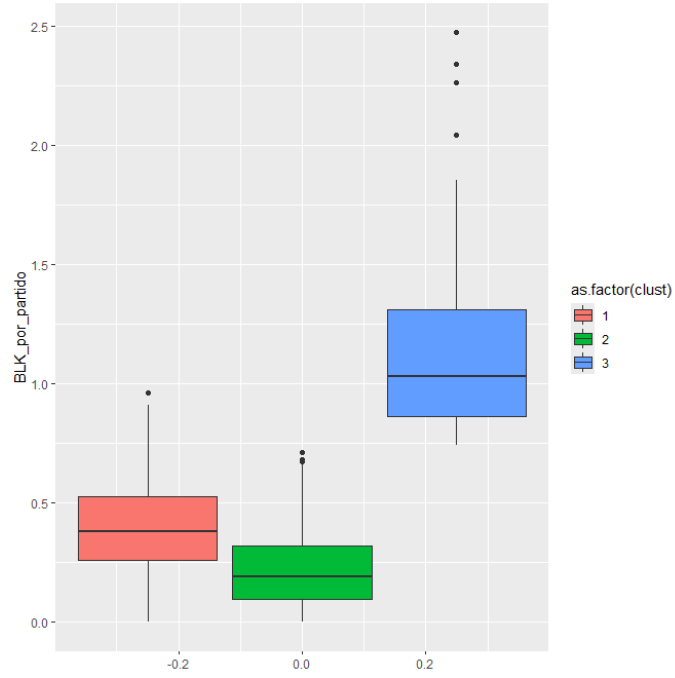


Figura 10: Distribución de tapones por partido según el clúster defensivo

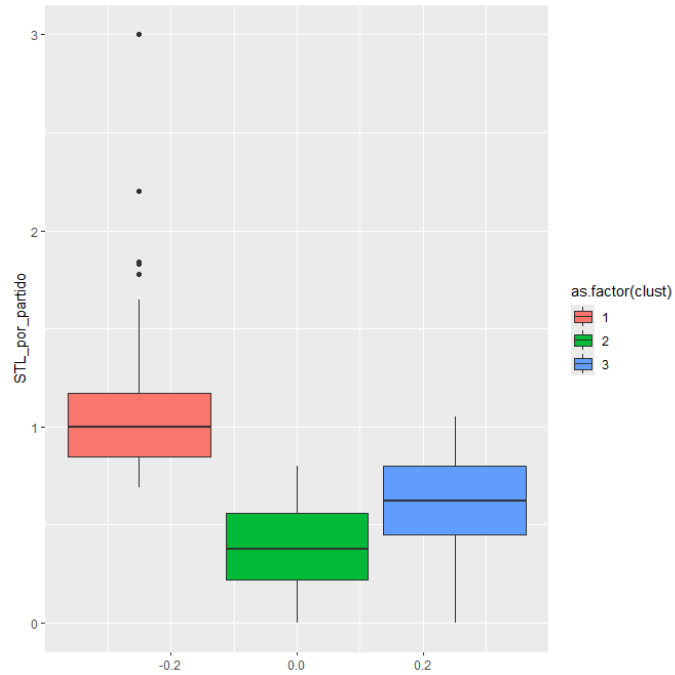


Figura 11: Distribución de robos por partido según el clúster defensivo

Posteriormente, exploramos la relación entre los grupos defensivos y la posición en el campo (variable *POS*). El test chi-cuadrado rechaza la hipótesis de independencia entre ambas variables ($p\text{-valor} < 2.2 \times 10^{-16}$), lo que indica que sí existe relación significativa. El gráfico de barras (Figura 12)

muestra cómo ciertas posiciones se asocian con tipos de defensa: por ejemplo, los pivots (C) se concentran más en el grupo de taponadores, mientras que los bases (PG) lo hacen en el de ladrones.

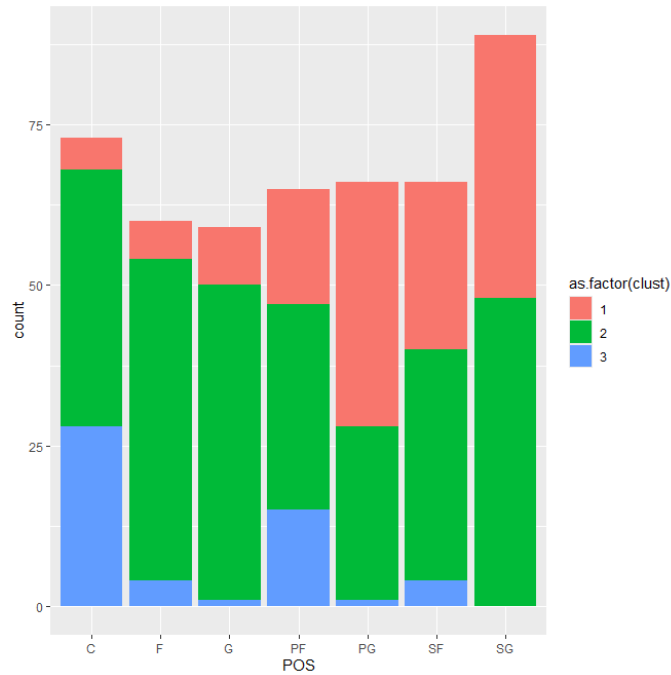


Figura 12: Distribución de posiciones por clúster defensivo

Para interpretar mejor esta relación, realizamos un análisis de correspondencias. En la Figura 13 se representa el plano de las dos primeras dimensiones. Se observa que los taponadores se asocian con posiciones interiores como pivots (C) y ala-pívots (PF), mientras que los ladrones se vinculan con la posición de base (PG). Los malos defensores, en cambio, no muestran una asociación clara con ninguna posición específica.

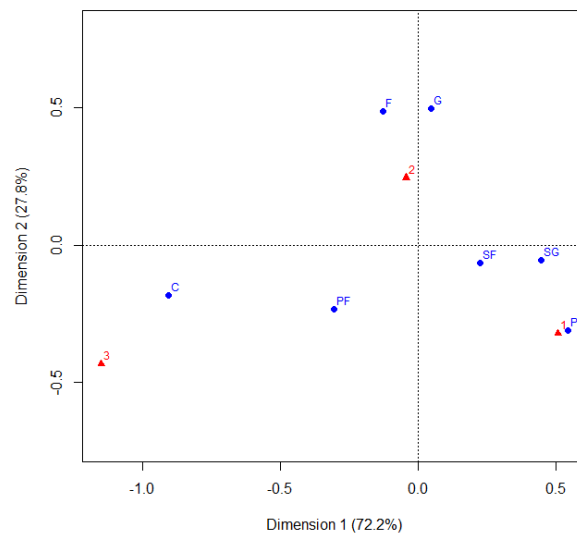


Figura 13: Análisis de correspondencias entre tipo de defensor y posición en el campo

Finalmente, analizamos si existen diferencias salariales entre los grupos defensivos. La Figura 14 muestra que los jugadores clasificados como ladrones tienden a tener los salarios más altos, seguidos por los taponadores, mientras que los malos defensores presentan los salarios más bajos.

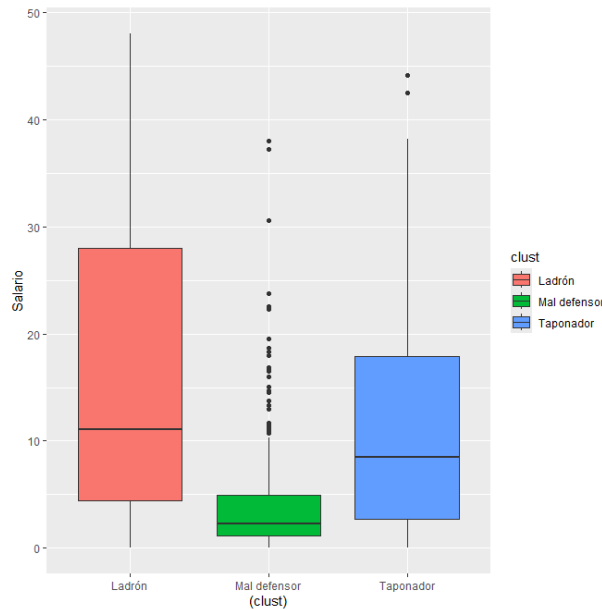


Figura 14: Distribución del salario según el tipo de defensor

El test de Kruskal-Wallis confirma que existen diferencias estadísticamente significativas entre los grupos defensivos ($p\text{-valor} < 2.2 \times 10^{-16}$) con respecto al salario. Para ver qué grupo o grupos estaban fallando, se utilizó el test de Wilcoxon con la corrección de Holm. Se obtuvo la siguiente tabla:

	Ladrón	Mal defensor
Mal defensor	0.000000000000000000003103023	NA
Taponador	0.1110940403098189671604600903	0.0000001491313442886075154182

Observamos diferencias significativas entre los ladrones y los malos defensores, así como entre los taponadores y los malos defensores. En cambio, no se encontraron diferencias significativas entre los ladrones y los taponadores.

2.5. Red neuronal para la predicción del salario

Dado que las herramientas de R resultaban poco eficientes para construir una red neuronal, optamos por utilizar Python a través de Google Colab, empleando la biblioteca `scikit-learn`. Nuestro objetivo era predecir el salario de los jugadores a partir de las variables que mostraban mayor relación con ésta en los análisis anteriores.

Construimos una red neuronal de tipo perceptrón multicapa (MLP), utilizando las variables previamente normalizadas. Empezamos entrenando el modelo, tomando de manera aleatoria un 80 % de la base de datos con las estadísticas de la temporada 22-23, y evaluamos su rendimiento en un conjunto de prueba independiente, es decir el 20 % restante.

En la Figura 15 mostramos la comparación entre el salario real y el predicho sobre los datos de prueba. La Figura 16 identifica algunos jugadores que presentan grandes discrepancias entre su salario real y el predicho, lo que podría interpretarse como casos de sobrevaloración o infravaloración. Analizando alguno de los casos con grandes discrepancias encontramos el caso de Myles Turner, el jugador durante la temporada 2022/23 cobró aproximadamente 35 millones de dolares, sin embargo la red neuronal le otorga un salario de aproximadamente 20 millones de euros, cantidad que ha cobrado las dos temporadas siguientes, tras su renovación. El caso del jugador Jonathan Isaac también es especial, podemos deducir que la red neuronal le otorga un salario tan bajo por haber jugado únicamente 11 partidos, es decir, estuvo lesionado gran parte de la temporada. Por el contrario, también nos encontramos con jugadores a los que la red neuronal les otorga un salario superior al real, casos como el del jugador Ja Morant. Estos últimos son casos de jugadores jóvenes, con pocos años en la liga y a los cuales les esperan renovaciones¹ mucho mejores.

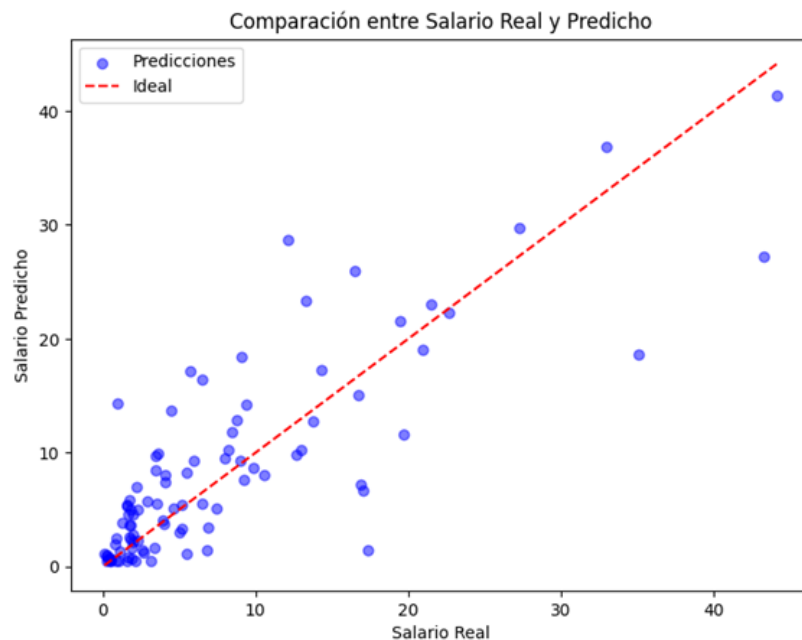


Figura 15: Comparación entre salario real y predicho (entrenamiento).

¹Ja Morant actualmente cobra 36 millones de dólares.

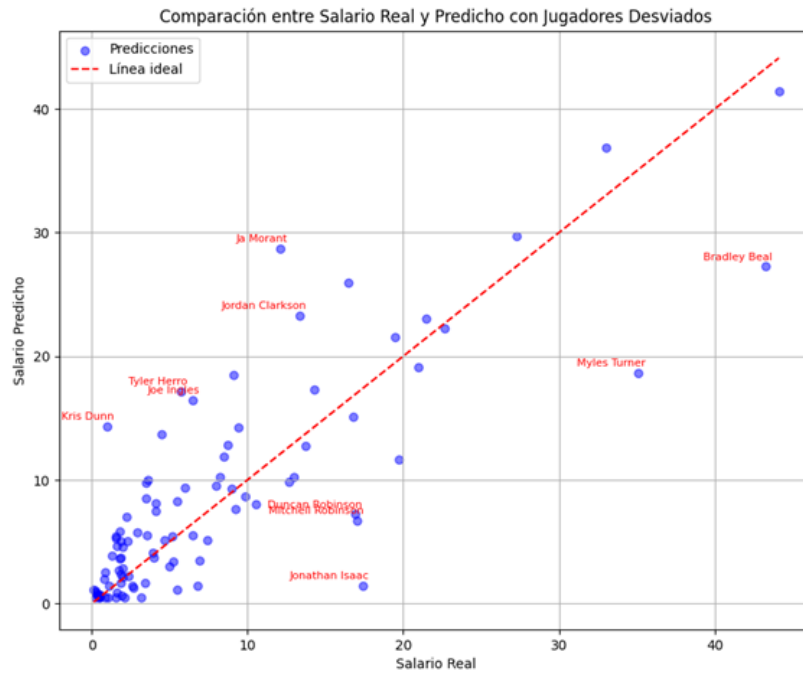


Figura 16: Jugadores con mayores desviaciones en el conjunto de entrenamiento.

Sin embargo, se ha considerado que el 20 % restante de la base de datos es un conjunto demasiado pequeño para analizar el comportamiento de la red neuronal, por ello se ha decidido tomar como nuevo conjunto de prueba las estadísticas de todos los jugadores activos de la temporada siguiente (la temporada 23-24). El comportamiento del modelo se muestra en la Figura 17, donde se observa una dispersión mayor. En la Figura 18 destacamos los diez jugadores con mayor diferencia absoluta entre el salario real y el estimado.

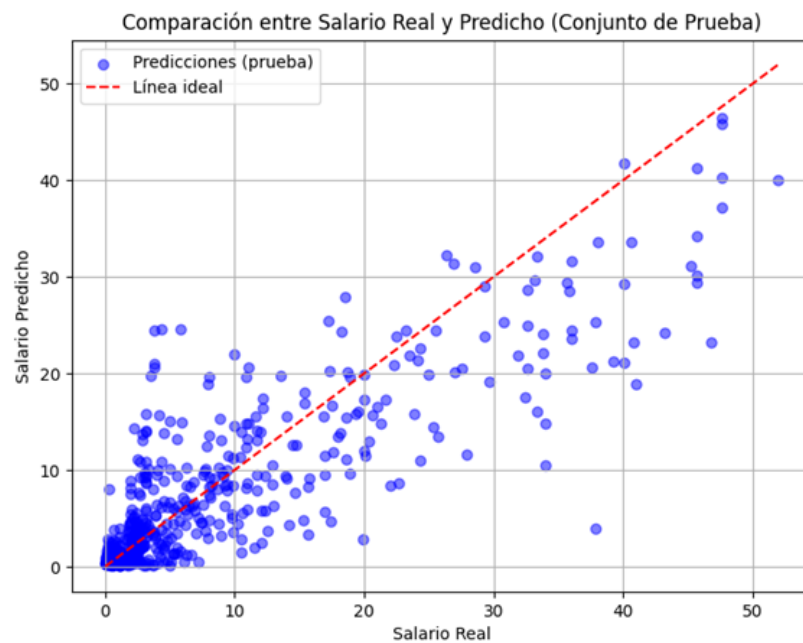


Figura 17: Comparación entre salario real y predicho (prueba).



Figura 18: Top 10 jugadores con mayor desviación en el conjunto de prueba.

Por último, analizamos la relación entre las predicciones y dos variables relevantes: la edad y los minutos jugados. En la Figura 19 mostramos la distribución de errores según tres grupos de edad: menores de 25 años, entre 25 y 30 años y mayores de 30 años. Y en la Figura 20 observamos cómo los errores se concentran especialmente entre los jugadores con pocos minutos disputados, donde la predicción resulta más difícil.

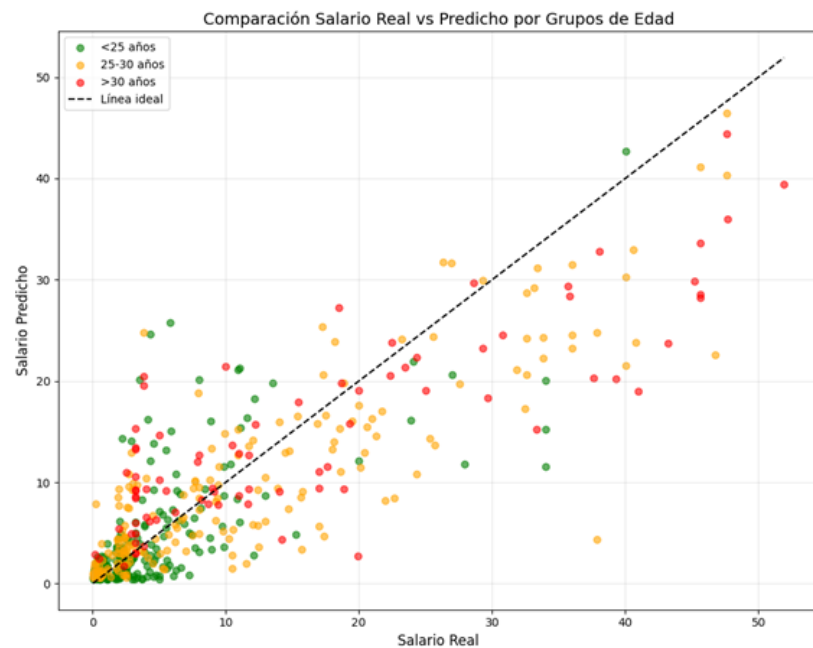


Figura 19: Salario real vs predicho por grupos de edad.

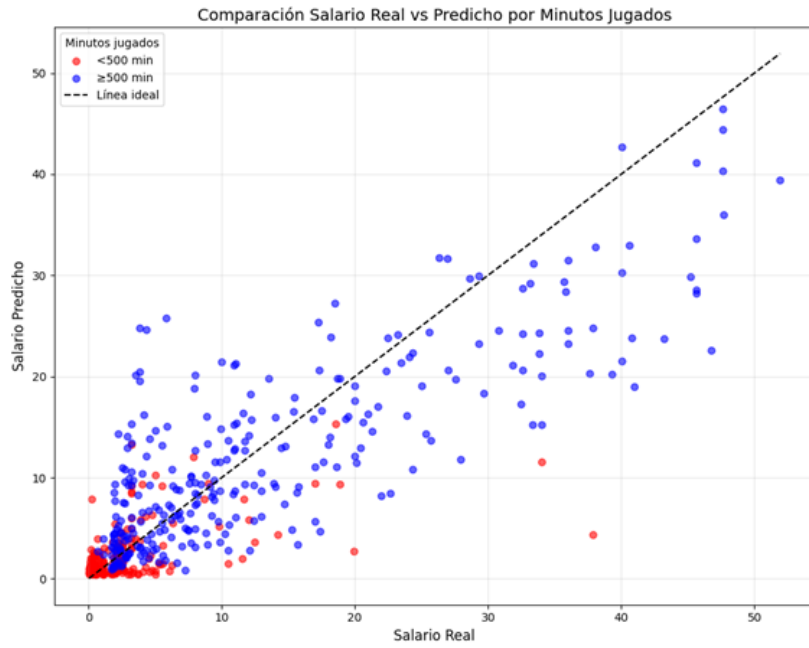


Figura 20: Salario real vs predicho por minutos jugados.

En conjunto, consideramos que la red neuronal logra capturar de forma razonable las relaciones entre el rendimiento y el salario de los jugadores, aunque presenta limitaciones en los extremos y en jugadores con escasa participación.

2.6. Relación entre la posición en el campo y el salario

En esta sección analizamos cómo se distribuyen los salarios en función de la posición que ocupa cada jugador en el campo. Para ello, representamos gráficamente el número de jugadores por grupo salarial y posición, como puede verse en la Figura 21.

Dado que las categorías F y G agrupan a jugadores que pueden desempeñarse en dos posiciones (por ejemplo, F combina SF y PF), decidimos eliminarlas del análisis para evitar ambigüedades y mejorar la fiabilidad estadística. Con este conjunto filtrado realizamos un test de independencia chi-cuadrado entre posición y grupo salarial. El resultado ($p\text{-valor} \approx 0.089$) no permite rechazar la hipótesis nula de independencia bajo un nivel de significación de $\alpha = 0.05$, por lo que no se encuentran evidencias suficientes para afirmar que la posición esté relacionada con el grupo salarial.

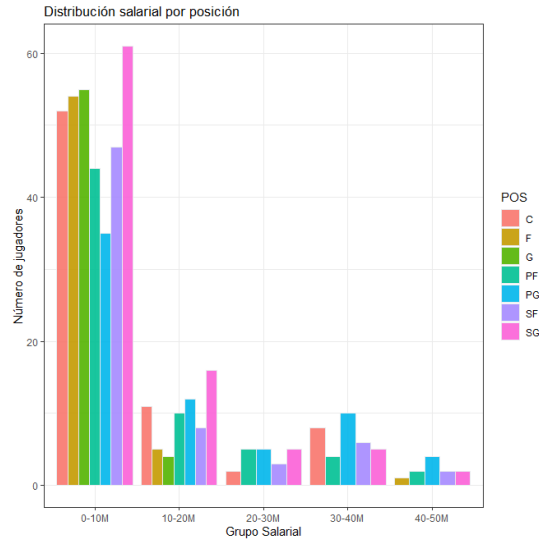


Figura 21: Distribución salarial por posición.

2.7. Análisis del salario por minuto

Adicionalmente, exploramos posibles casos atípicos evaluando el salario por minuto jugado frente a la eficiencia (Figura 22). Esta visualización nos permitió identificar a jugadores con muy pocos minutos jugados pero salarios elevados, lo que probablemente se debe a lesiones de larga duración. Con esta información, filtramos a los jugadores con ratios de salario por minuto excesivamente altos (mayores a 0,1 millones de dólares por minuto jugado) y preparamos un nuevo conjunto de datos, con el que trabajaremos en la próxima sección.

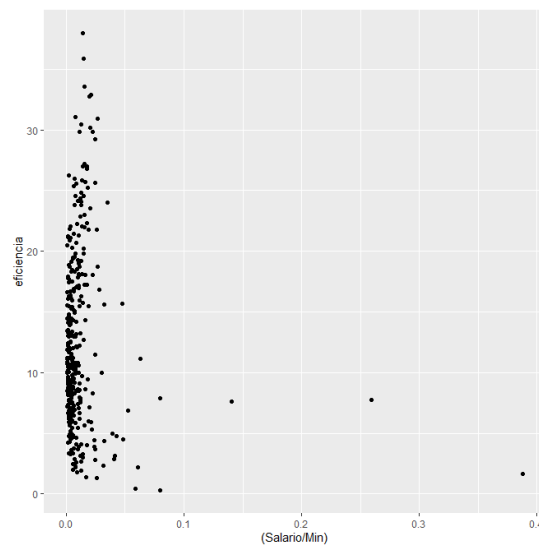


Figura 22: Relación entre eficiencia y salario por minuto jugado

2.8. Relación entre el perfil de tiro y el salario

El baloncesto actual destaca por su elevado número de triples intentados por partido, es por ello que en un primer momento, nos preguntamos si existía una relación directa entre el salario de los jugadores y su capacidad triplista. Para explorarlo, comenzamos analizando la relación entre el salario y el número de triples anotados ($X3PM$) por posición. La Figura 23 muestra la distribución de ambos valores. Aunque algunas posiciones, como la de escolta (SG), presentan una correlación moderada ($r \approx 0.6$), no encontramos una relación fuerte entre estas dos variables.

Para ampliar este análisis, decidimos sintetizar la información relacionada con el tiro (tiros intentados y anotados de campo, triples y tiros libres) mediante un análisis de componentes principales (PCA).

La Figura 24 representa la varianza explicada por cada componente. Observamos que la primera componente principal explica un 79.6 % de la variabilidad, y que las dos primeras componentes en conjunto superan el 97 %.

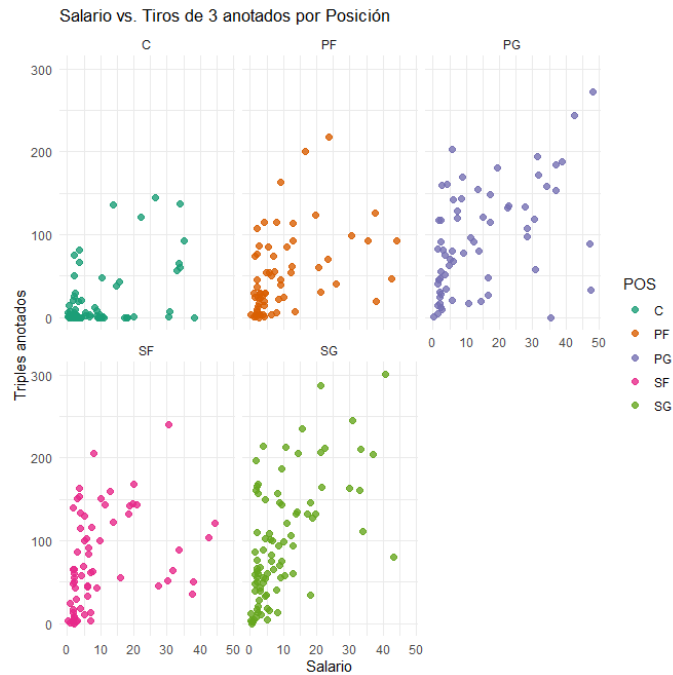


Figura 23: Salario vs. triples anotados ($X3PM$) por posición.

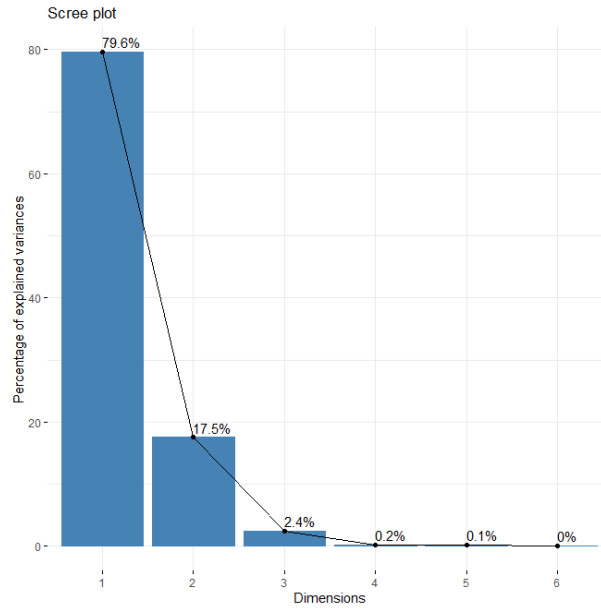


Figura 24: Varianza explicada por las componentes principales

En la Figura 25 representamos las variables originales de tiro en el plano generado por las dos primeras componentes principales. Observamos, por el color de las flechas, que las variables de tiro están muy bien explicadas por este nuevo sistema.

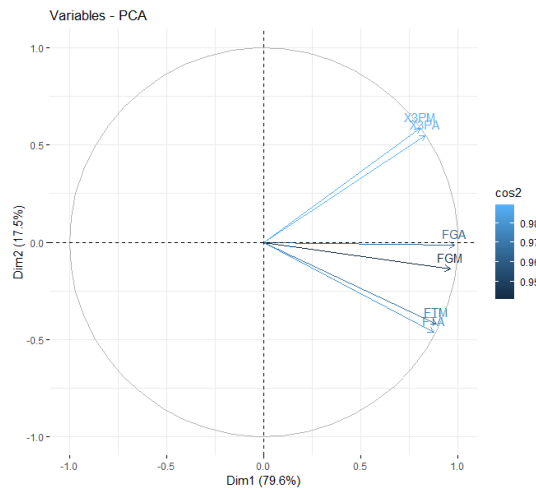


Figura 25: Representación de las variables de tiro en el plano principal del PCA

La primera componente recoge en gran medida la capacidad anotadora del jugador, tanto por el volumen de lanzamiento como por su nivel de acierto. Al calcular su correlación con el salario, obtenemos un valor de $r = 0.64$. En la Figura 26, visualizamos los individuos (jugadores) proyectados en el plano de las componentes principales, coloreados según su salario.

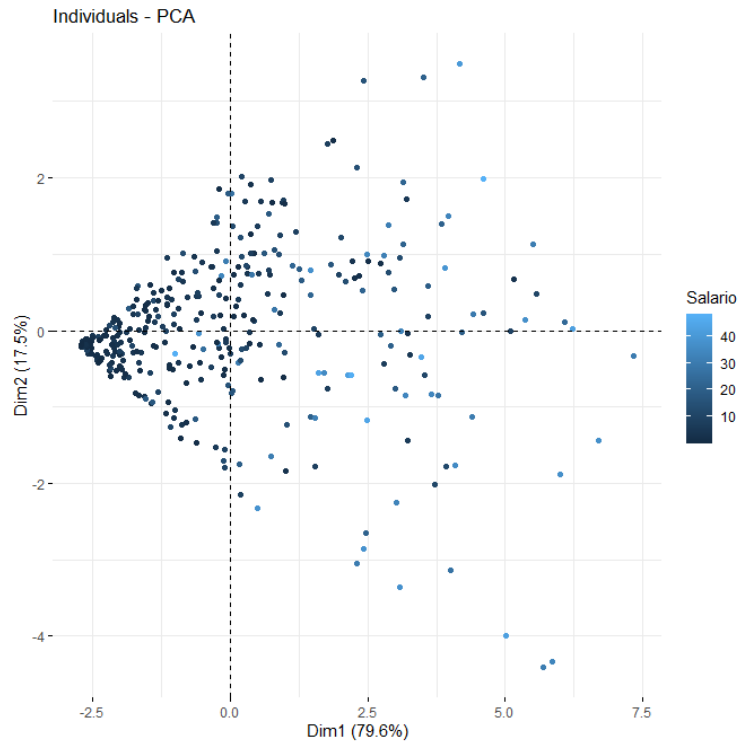


Figura 26: Individuos proyectados en el espacio PCA coloreados por salario

En la Figura 27, representamos la relación entre la primera componente y el salario, segmentando por posición. Llama especialmente la atención el caso de los pívots (C), donde la correlación supera el 70 %.



Figura 27: Salario vs. primera componente principal por posición

Estos resultados nos llevaron a centrar el análisis en la evolución de la posición de pívot. Históricamente, los jugadores interiores (pívot(C) y ala-pívot(PF)) se centraban en el juego cerca del aro. Sin embargo, a partir de los años 2000, algunos pívots y ala-pívots comenzaron a ampliar su rango de tiro. Jugadores como Dirk Nowitzki revolucionaron el baloncesto con su habilidad para anotar desde cualquier parte del campo, abriendo el juego y cambiando el perfil ofensivo de los jugadores altos. Hoy en día, se valora enormemente a los pívots con amenaza exterior, ya que permiten ofensivas más abiertas y dinámicas.

En base a esta transformación del juego, decidimos ajustar un modelo de regresión lineal entre la primera componente principal y el salario, centrándonos en los pívots. La Figura 28 ilustra este ajuste, con un coeficiente de determinación de $R^2 \approx 0.55$, lo que refleja una relación considerable.

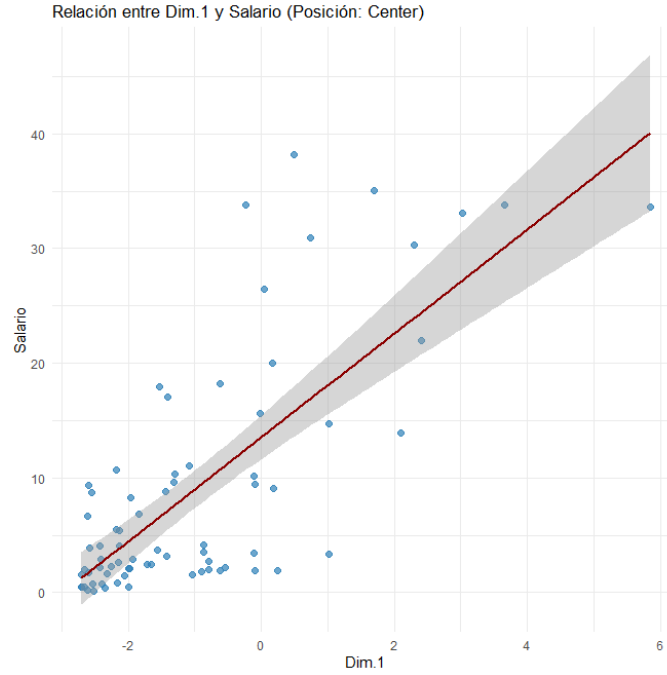


Figura 28: Relación entre la primera componente y salario en la posición de pívot

Para mejorar esta predicción, decidimos construir un árbol de regresión utilizando como variables explicativas las dos primeras componentes y la eficiencia. Esta última variable ha demostrado ser muy relevante para identificar jugadores de alto rendimiento. En la Figura 29, se representa el árbol resultante. *Dim.1* representa la primera componente principal y *Dim.2* la segunda componente principal.

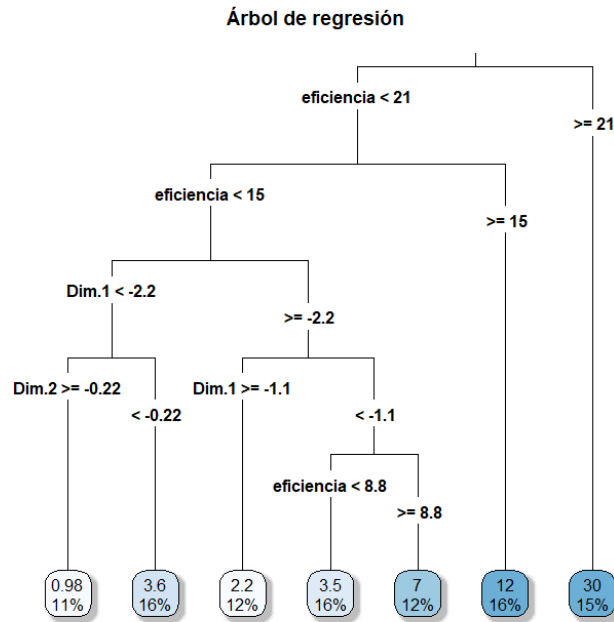


Figura 29: Árbol de regresión para predecir el salario en la posición de pívot.

La Figura 30 muestra las predicciones del árbol frente a los salarios reales. El error cuadrático medio (RMSE) fue de 4.19 millones, lo que consideramos razonable dadas las diferencias individuales entre jugadores.

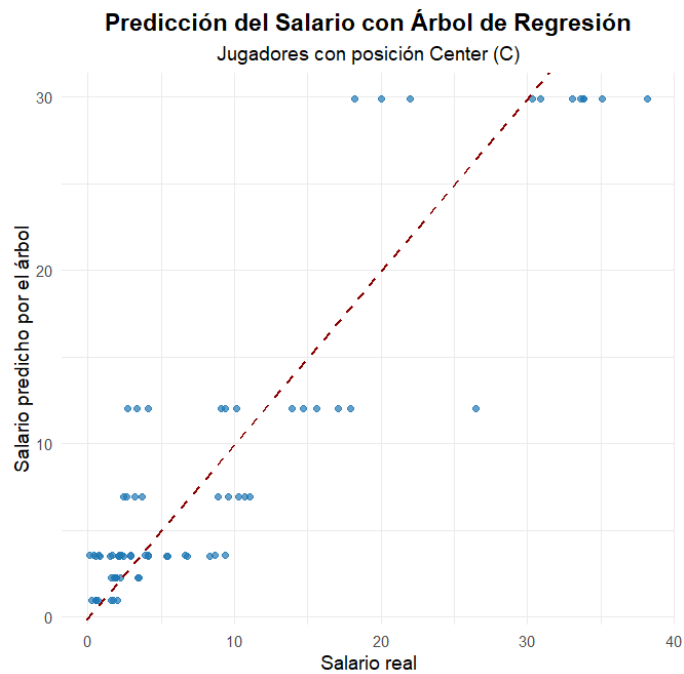


Figura 30: Salario real vs. predicho por el árbol de regresión (pívots).

En definitiva, este análisis evidencia que el perfil ofensivo global, capturado por la primera com-

ponente del PCA, es un buen predictor del salario, especialmente en la posición de pívot. La incorporación de la eficiencia permite afinar aún más las predicciones, destacando los casos de jugadores muy productivos en distintos roles ofensivos.

3. Conclusiones

Tras todo el trabajo de análisis e interpretación de los datos estadísticos desarrollado, hemos sacado en claro las siguientes reflexiones:

Hemos podido comprobar, tal y como intuíamos desde un principio, que los salarios de los jugadores de la liga guardan cierta relación con su rendimiento en la pista. Sin embargo, con nuestros análisis y datos manejados, no somos capaces de hacer una predicción exacta de dicha retribución. Factores extra a las estadísticas tienen una influencia notable, en el deporte, no todo se mide con números.

Un ejemplo claro es la dinámica del mercado: cuando varios equipos se disputan a un jugador estrella, este suele decantarse por quien le ofrezca el mejor contrato. Aunque hay excepciones, como jugadores que priorizan un proyecto deportivo antes que el dinero, lo habitual es que el aspecto económico pese más.

También puede suceder que, tras la firma de un contrato nuevo, el jugador en cuestión baje su rendimiento. Con el dinero y plaza en el equipo asegurados, algunos se relajan y bajan el nivel en la pista, y eso se refleja en nuestros datos. Nuestro análisis detecta esos jugadores que, según sus estadísticas, deberían recibir menos dinero del que les firmaron sus respectivos equipos.

La mala fortuna en el mundo del deporte muchas veces juega un papel importante, y las lesiones seguramente sean el principal foco de esta. Como pudimos suponer haciendo los análisis, hay jugadores con contratos muy importantes, que no llegan a jugar muchos partidos por problemas de salud, alterando las predicciones de rendimiento estadístico en función del salario pactado en sus contratos.

Siguiendo en la línea de buscar el porqué de aquellos que según nuestro análisis son sobrepagados, podemos mencionar situaciones de jugadores considerados como leyendas del club, las cuales por el rendimiento del pasado mantienen una buena renta económica, pese a su bajón inevitable de nivel, por una cuestión natural: el paso del tiempo. Lo normal es que asuman un papel más de mentores, enseñando y corrigiendo a los más jóvenes. Y aunque algunos de estos experimentados jugadores se rebajan el salario voluntariamente para ayudar al equipo a reforzarse, no es lo más común.

Por otra parte, no hemos detectado que se premie lo suficiente económicamente a los mejores defensores, y eso que en el baloncesto es tan importante sumar puntos, como evitar que los metan. Gana el que anota más puntos que el rival, sin importar el número total de estos. Podemos suponer que esto tiene que ver con el espectáculo, ya que muchas veces el espectador le da mayor importancia al que más canastas mete, o al que hizo el mate más impresionante, o el triple desde más lejos, dejando para un segundo plano a aquel que molestó e impidió que anotase la estrella rival, permitiendo a su

equipo ganar el partido.

También concluimos que la posición no afecta al salario de los jugadores. Desde cada una de ellas, observamos cómo los mejores estadísticamente obtienen los mejores contratos, indistintamente de su lugar en el campo. Al final, las franquicias pagan mejor a los que mejor rinden, y en el baloncesto puedes rendir el que más desde cualquiera de las posiciones. Particularizando a la temporada observada, la discusión del jugador más valorado (MVP), premio que se otorga finalizada la temporada regular, mediante votación entre la prensa especializada de la liga, se centró en dos jugadores, Nikola Jokic y Joel Embiid, ocupando estos la posición de pívot, algo que a lo largo de la historia moderna de la NBA fue poco habitual.

En cuanto al análisis realizado mediante redes neuronales, hemos podido detectar que los jugadores veteranos son los que pueden acceder, y de hecho acceden, a mejores contratos. Esto se debe a más restricciones y control que hace la dirección de la liga sobre los contratos de los jugadores. Para los recién llegados, conocidos como *rookies*, el límite superior de retribución es menor que para aquellos jugadores que tengan más años cotizados en la liga.

Visualizando nuestro trabajo realizado, nos dimos cuenta de por qué se usa tanto en el análisis de datos baloncestístico la variable *eficiencia*, una combinación que recoge los principales atributos estadísticos, tanto positivos como negativos, mediante una fórmula, para medir de alguna forma numérica, la contribución de los diferentes integrantes del equipo en los partidos. Esta variable es la que mejor nos ha ayudado a percibir cuánto cobra un jugador actualmente, o puede llegar a hacerlo en su próximo contrato, de todas las que hemos manejado en nuestro informe.

4. Reparto de tareas

1. **Búsqueda de los datos:** David Pérez Navarro
2. **Realización del código:** David Pérez Navarro, Cristina Menéndez Fernández
3. **Realización de la introducción:** David De Soto Manilla
4. **Redacción de los resultados obtenidos a partir del código:** Pablo Llorián González
5. **Redacción de las conclusiones:** David Pérez Navarro, David De Soto Manilla, Cristina Menéndez Fernández
6. **Corrección del informe:** Cristina Menéndez Fernández, David Pérez Navarro
7. **Desarrollo de la presentación:** Cristina Menéndez Fernández, Yaiza Martínez Jiménez
8. **Revisión del informe:** Yaiza Martínez Jiménez

Para la realización de este trabajo se utilizó la herramienta de IA, *chatgpt*, para:

- Búsqueda de comandos para la representación gráfica de los resultados en *R*
- Funcionamiento de *Python* a través de *Google Colab*

5. Bibliografía

- Apuntes de la asignatura *Análisis de Datos* de la Universidad de Oviedo
- https://www.basketball-reference.com/leagues/NBA_2023_totals.html, donde se reflejan las estadísticas de los jugadores NBA durante la temporada 2022/2023.
- <https://hoopshype.com/salaries/players/2022-2023/>
- <https://openai.com/chatgpt/overview/>
- <https://fhernanb.github.io/Manual-de-R/>

A. Código

Script de análisis en R

```
# DATOS ----  
#vamos a llamar nba a la copia de datos_combinados  
datos_combinados<-read.csv("C:/Users/david/Downloads/datos_combinados.csv")  
datos_combinados<-read.csv("C:/Users/cris/Desktop/datos_combinados.csv")  
nba<-datos_combinados  
#La base de datos cuanta con NAs  
  
which((nba$POS=="N/A"))  
  
head(nba[which((nba$POS=="N/A")),]) #son 4 jugadores con 1 a 4 partidos en la NBA  
#y no llegan a los 6 minutos por partido, podemos eliminarlos.  
  
nba <- nba[-c(14, 112, 142, 427), ]  
  
#Vamos a redefinir la variable salario representandola en millones de euros  
  
nba$Salario<-as.numeric(nba$Salario/1000000)  
#Y creamos una división por grupos del salario  
nba$Grupo_Salario <- cut(nba$Salario,  
                        breaks = seq(0, 50, by = 10),  
                        labels = c("0-10M", "10-20M", "20-30M", "30-40M", "40-50M"),  
                        include.lowest = TRUE)  
  
#Es interesante añadir las medias por partido de los jugadores  
nba <- nba %>%  
  mutate(  
    PTS_por_partido = round(PTS / GP,2),  
    AST_por_partido = round(AST / GP,2),  
    REB_por_partido = round(REB / GP,2),  
    BLK_por_partido = round(BLK / GP,2),  
    STL_por_partido = round(STL / GP,2)  
  )  
  
#Además en estadística avanzada de la NBA se utiliza la fórmula de eficiencia  
nba$eficiencia<- (nba$PTS + nba$REB+ nba$AST + nba$STL + nba$BLK - (nba$FGA -nba$FGM)  
  ↪ -(nba$FTA -nba$FTM) -nba$TOV)/nba$GP
```

```

# COMPONENTES PRINCIPALES ----
library(tidyverse)
library(FactoMineR)
library(factoextra)
library(corrplot)
# Eliminar 'Salario' para este primer PCA y las variables no numéricas
datos_numericos <- nba %>%
  select_if(is.numeric)
rendimiento <- datos_numericos %>% select(-Salario)

# PCA con centrado y escalado
pca1 <- PCA(rendimiento, scale.unit = TRUE, #esto normaliza los datos (media 0, varianza
  ↪ 1), muy importante si las variables están en diferentes escalas (por ejemplo, puntos y
  ↪ minutos).
  graph = FALSE)
summary(pca1)
#pca1$var$coord
#pca1$var$cos2
#pca1$var$contrib #contribucion de cada variable a la componente

### Dim.1 ----
#Es un claro componente de rendimiento total y volumen de juego.
#Resume a los jugadores que:
##Juegan muchos minutos.
##Anotan mucho.
##Toman rebotes y dan asistencias.
##Tienen alta eficiencia y producción global (fantasy points, etc.).
#Esta componente separa a los jugadores titulares con alto impacto del resto.

### Dim.2 ----
#Parece separar a:
##Jugadores interiores o defensivos (más rebotes ofensivos y bloqueos, mejor FG%), de
##Jugadores perimetrales que tiran más triples (por eso esas variables tienen cargas
  ↪ negativas).
#Esta dimensión puede llamarse "estilo de juego: interior vs. exterior".

### Dim.3 ----
#Está más relacionada con jugadores versátiles o completos, que:
##Consiguen dobles-dobles / triples-dobles.
##Tienen buenas asistencias por partido.
##Son jugadores "all-around".

```

#Esta dimensión puede interpretarse como “versatilidad estadística” o “impacto completo en el juego”.

Visualizar resultados

```
fviz_eig(pca1, addlabels = TRUE) # varianza explicada
```

fviz_pca_var(pca1, col.var = "cos2") # este gráfico nos dice como de bien estan representadas las variables con las dos primera componentes, cuanto más larga sea la flecha mejor explicada estara.

#Además, la direccion de la flecha nos indica qué componente determina esa variable #si está sobre el eje horizontal está directamente relacionada con la primera #y analogamente si está en vertical con la segunda.Si está entre ambas explica las dos. #Por ultimo, el angulo entre las flechas nos dice cómo estan relacionadas las variables. #Un angulo de 90 grados indica que no están nada relacionadas, #flechas cercanas y con ángulo similar, están positivamente correladas, #y flechas que formen un ángulo de 180 grados están negativamente correladas.

Relación del salario con componentes principales

```
cor(datos_numericos$Salario, pca1$ind$coord[,1]) # Correlación con Dim 1
```

```
cor(datos_numericos$Salario, pca1$ind$coord[,2]) # Correlación con Dim 2
```

```
cor(datos_numericos$Salario, pca1$ind$coord[,3]) # Correlación con Dim 3
```

#Conclusion:

#Salario está más influido por el rendimiento total (PC1) que por estilo de juego o versatilidad.

#PC1 podría servir como una variable resumida para predecir o explicar el salario.

#PC3 también aporta al salario, así que un modelo con PC1 y PC3 podría ser útil.

Añadir salario a scatterplot de jugadores

```
fviz_pca_ind(pca1,
             geom.ind = "point",
             col.ind = datos_numericos$Salario,
             palette = "viridis",
             addEllipses = FALSE,
             legend.title = "Salario")
```

#En este gráfico vemos lo que comentamos antes, la dimensión 1 determina bastante

#el salario, es decir, a mayor componente principal 1, mejor es el salario

#Vamos a visualizar las primeras tres componentes:

```
library(plotly)
```

Coordenadas de las 3 primeras componentes

```
coords <- pca1$ind$coord[, 1:3]
```

```

# Gráfico 3D con color por salario
plot_ly(x = ~coords[,1], y = ~coords[,2], z = ~coords[,3],
        type = "scatter3d", mode = "markers",
        marker = list(size = 4,
                      color = datos_numericos$Salario, # color por salario
                      colorscale = "Viridis",
                      showscale = TRUE),
        text = ~paste("Salario: ", datos_numericos$Salario)) %>%
layout(scene = list(
  xaxis = list(title = "PC1"),
  yaxis = list(title = "PC2"),
  zaxis = list(title = "PC3")),
  title = "Jugadores: Componentes principales vs Salario")
#De nuevo vemos lo comprobado anteriormente, además de ver como la tercera componente
#también influye, aunque en menor medida, en el salario

# SALARIO ----

## Representación grafica ----
ggplot(nba,aes(x=Grupo_Salario))+ geom_bar( fill="#1874CD", color="#e9ecef",alpha=0.9)+
  theme_bw()

#Vamos ahora a relacionar la eficiencia de los jugadores con su salario

## Jugadores clave ----
jugador_mas_eficiente <- nba %>% arrange(desc(eficiencia)) %>% slice(1)
jugador_mas_caro <- nba %>% arrange(desc(Salario)) %>% slice(1)

ggplot(nba, aes(x = Salario, y = eficiencia)) +
  geom_point(color = "gray50", alpha = 0.6) + # Puntos normales en gris
  geom_point(data = jugador_mas_eficiente,
            color = "red", size = 3) + # Punto rojo para más eficiente
  geom_point(data = jugador_mas_caro,
            color = "red", size = 3) + # Punto rojo para más caro
  geom_text(data = jugador_mas_eficiente,
            aes(label = Jugador),
            color = "red", hjust = -0.1, vjust = 0.5) + # Nombre más eficiente
  geom_text(data = jugador_mas_caro,
            aes(label = Jugador),
            color = "red", hjust = -0.1, vjust = 0.5) + # Nombre más caro

```

```

labs(title = "Relación entre Salario y Eficiencia en la NBA",
      x = "Salario (en millones)",
      y = "Índice de Eficiencia",
      caption = "Jugadores más destacados en rojo") +
theme_minimal()

ggplot(nba,aes(y=eficiencia,fill=Grupo_Salario))+
  geom_boxplot()

#Veamos si estas dos variables guardan alguna relacion.
#Vamos a trabajar con la variable Salario, en vez de Grupo_Salario
#para aplicar un test de correlacion de Pearson

## Correlación de Pearson ----
cor.test(nba$eficiencia, nba$Salario) #p.valor<2.2e-16
#Luego como era de esperar, rechazamos la hipótesis nula, las variables
#Salario y eficiencia no son independientes

# CLUSTER ----
#cluster sobre robos por partido y tapones por partido (estadísticas defensivas), agrupa
↪ en malos defensores buenos bloqueadores
#y buenos ladrones
#Análisis usando paquetes de R

library(factoextra)
library(cluster)
library(tidyverse)
library(dplyr)

# La función select se mezcla con otra librería, hay q especificar que queremos usar
# la de dplyr
datos_cluster <- nba %>%
  dplyr::select(BLK_por_partido, STL_por_partido) %>%
  scale()

# K-means con visualización mejorada
set.seed(123)
km_res <- kmeans(datos_cluster, centers = 3, nstart = 25)

# Visualización
fviz_cluster(km_res, data = datos_cluster,

```

```

    ellipse.type = "convex",
    repel = TRUE,
    ggtheme = theme_bw(),
    geom="point") +
labs(title = "Clustering de Jugadores por Bloqueos y Robos",
     subtitle = "3 grupos identificados") +
scale_color_brewer(palette = "Set1") +
scale_fill_brewer(palette = "Set1")

nba2<-nba
nba2$clust<-km_res$cluster

ggplot(nba2,aes(y=BLK_por_partido,fill=as.factor(clust))) +
  geom_boxplot() #el grupo 3 es un grupo con grandes taponadores

ggplot(nba2,aes(y=STL_por_partido,fill=as.factor(clust))) +
  geom_boxplot() #el grupo 1 es un grupo con grandes "ladrones"

#existe alguna relación entre el grupo y la posición en la que juegan?

ggplot(nba2,aes(x=POS,fill=as.factor(clust))) +
  geom_bar()

chisq.test(table(nba2$POS,nba2$clust))
#Rechazamos la hipótesis nula de independencia bajo un nivel de
#significación de 0.05, si que existe relación entre posición
#y tipo de defensor. Veamos como es ese tipo de relación mediante un
#análisis de correspondencia

install.packages("ca")
library(ca)
A<-table(nba2$POS,nba2$clust)
sal<-ca(A)
resumen<-summary(sal)
resumen
plot(sal)
#entre los grandes taponadores
# de la liga destacan posiciones interiores c y pf
#el grupo de jugadores con más robos destaca la posición de base PG

```

```

nba2[,39]      #la variable que hace referencia al grupo al que pertenece
#1==ladrón, 2==mal defensor, 3==taponadores

#le asignamos un nombre a los valores de la variable clust
#el siguiente comando ejecutar únicamente una vez, si no se producen NAs
nba2$clust <- factor(nba2$clust,
                     levels = c(1, 2, 3),
                     labels = c("Ladrón", "Mal defensor", "Taponador"))

ggplot(nba2,aes(y=Salario,x=(clust),fill=clust)) + geom_boxplot()

summarise(group_by(nba2,clust),
           Media_partidos=mean(GP),
           DesvTip_partidos=sd(GP),
           numero=length(GP),
           media_salario=mean(Salario),
           min_partido_media=mean(Min/GP),
           min_partido_sd=sd(Min/GP)
)

head(nba2[nba2$clust == "Mal defensor", c(1:10)][order(nba2$Salario[nba2$clust == "Mal
↪ defensor"], decreasing = TRUE), ])

kruskal.test(Salario ~ clust, data=nba2)

#Rechazamos la hipótesis nula (Al menos una muestra proviene
#de una población con distribución distinta) para un nivel de significación
#de 0.05 (y a cualquier nivel de significación en realidad). Veamos que grupos
#difieren

wilcox_post<-pairwise.wilcox.test(x=nba2$Salario,g=nba2$clust,p.adjust.method = "holm")
format(wilcox_post$p.value, scientific = FALSE) #representación sin notación científica
#para ver mejor como de pequeño es el p-valor.

#Existe una diferencia evidente entre el grupo de malos defensores y el de taponadores

```

#Existe una diferencia evidente entre el grupo de malos defensores y el de ladrones

#Hemos encontrado grupos con diferencias significativos en el Salario

#RED NEURONAL----

#

*# #Debido a la ineficiencia de las librerías de R para crear una red neuronal, se ha
decidido recurrir a python, mediante google colab, más concretamente hemos decidido
trabajar con la librería sk-learn para desarrollar así una red neuronal que a partir
de las variables con mas relación con el Salario para predecirlo. A la hora de
ejecutar el código del google colab es necesario el excel nba2.csv y prueba_colab2.*

Respecto a la red neuronal se comenta el uso de Google Colab para su implementación, dicho código se encuentra en [Red_Neuronal](#).

POSICION ----

```
library(ggplot2)
```

```
library(dplyr)
```

Representacion grafica por POS ----

```
ggplot(nba, aes(x = Grupo_Salario, fill = POS)) +  
  geom_bar(position = "dodge", color = "#e9ecef", alpha = 0.9) +  
  labs(title = "Distribución salarial por posición",  
        x = "Grupo Salarial",  
        y = "Número de jugadores") +  
  theme_bw()
```

#Veamos la relación del grupo salarial con la posición en la que juegan,

#quitamos las posiciones "F" y "G" porque son los que pueden jugar como dos posiciones

#son pocos jugadores y afectan a los resultados

```
chisq.test(table(nba$Grupo_Salario[which(nba$POS!=c("F","G"))],nba$POS[which(nba$POS!=c("F","G"))])
```

#obtenemos un p.valor menos el que el nivel de significacion alfa=0.05,

#luego rechazamos la hipótesis nula, se tiene que la posición y el salario son variables

↪ independientes

#Por el comentario anterior, eliminamos a los jugadores que juegan en la posición "F",

↪ mezcla

#de SF Y PF, y G mezcla de PG Y SG.

```
nba2_filtrado <- nba2 %>%  
  filter(!POS %in% c("F", "G"))
```

```

#Detecto estos jugadores que tienen un salario altísimo por minuto jugado,
#es decir jugadores con lesiones de larga duración

ggplot(nba2_filtrado,aes(x=(Salario/Min),y=eficiencia))+
  geom_point()

nba2_filtrado<-nba2_filtrado[-which((nba2_filtrado$Salario/nba2_filtrado$Min)>=0.1),]

descarga2<-nba2_filtrado[,c(1,2,5,9,10,11,12,14,15,17,18,21,23,24,25,33,34,35,38)]

#Guardo el archivo descarga2 en la memoria del ordenador

write.csv(descarga2, file = "descarga2.csv", row.names = FALSE)

## Tiro ----
#En un principio no parece existir una relación fuerte entre el salario
#y los triplistas

summarise(group_by(nba2_filtrado,POS),
  correlacion=cor(Salario,X3PM))
ggplot(data = nba2_filtrado, aes(x = Salario, y = X3PM, colour = POS)) +
  geom_point(alpha = 0.8, size = 2) +
  scale_color_brewer(palette = "Dark2") +
  facet_wrap(~ POS) +
  labs(title = "Salario vs. Tiros de 3 anotados por Posición",
    x = "Salario",
    y = "Triples anotados") +
  theme_minimal()

#No parece existir una relación fuerte, quizás más en SG que el resto de posiciones
#Voy a transformar todas las variables de tiro en componentes principales

library(tidyverse)
library(FactoMineR)
library(factoextra)
library(corrplot)

#Creo la variable tiro, formada por tiros intentados y anotados, triples intentados
#y anotados, y tiros libres intentados y anotados. Hago componentes principales

```

```

#con estas variables

tiro<-nba2_filtrado[,c(11,12,14,15,17,18)]
pca1 <- PCA(tiro, scale.unit = TRUE,
            graph = FALSE)
summary(pca1)

# Visualizar resultados
fviz_eig(pca1, addlabels = TRUE) # varianza explicada
fviz_pca_var(pca1, col.var = "cos2") # este gráfico nos dice cómo de bien están
#representadas las variables con las dos primeras componentes, cuanto más larga sea
#la flecha mejor explicada estara.

#Con las dos primeras componentes explico la mayoría de la variabilidad

# Relación del salario con componentes principales
cor(nba2_filtrado$Salario, pca1$ind$coord[,1]) # Correlación con Dim 1
cor(nba2_filtrado$Salario, pca1$ind$coord[,2]) # Correlación con Dim 2

# Añadir salario a scatterplot de jugadores
fviz_pca_ind(pca1,
             geom.ind = "point",
             col.ind = nba2_filtrado$Salario,
             palette = "viridis",
             addEllipses = FALSE,
             legend.title = "Salario")

#Añado estos valores a la base de datos Coordenadas de las 2 primeras componentes
coords <-pca1$ind$coord[, 1:2]
nba2_filtrado<-cbind(nba2_filtrado,coords)
ggplot(nba2_filtrado,aes(x=Dim.1,y=Dim.2,colour=Salario))+geom_point()

summarise(group_by(nba2_filtrado,POS),
           correlacion1=cor(Salario,Dim.1),
           correlacion2=cor(Salario,Dim.2))

```

*#La primera componente principal tiene más correlación con el salario
#que la probada anteriormente de triples anotados*

```
ggplot(data = nba2_filtrado, aes(y= Salario, x = Dim.1, colour = POS)) +  
  geom_point(alpha = 0.8, size = 2) +  
  scale_color_brewer(palette = "Dark2") +  
  facet_wrap(~ POS) + theme_minimal()  
#
```

Evolucion del Pivot ----

*# # Resulta muy interesante observar la gran correlación entre la primera componente
principal y la posición de pivot. A partir de aquí iniciaremos una nueva parte del
estudio.*

*# # La posición de pivot en la NBA ha evolucionado notablemente en las últimas décadas.
En los años 80, era habitual que tanto el ala-pivot (PF) como el pivot (C) se
→ desempeñaran*

*# principalmente en zonas interiores, cerca del aro. Sin embargo, con la llegada de los
→ años 2000,
comenzó un cambio: los ala-pívots empezaron a ampliar su rango de acción hacia el
→ perímetro.*

*# Surgieron grandes ala-pívots con capacidad de anotar desde cualquier zona ofensiva del
campo, como Dirk Nowitzki, quien marcó una revolución en el perfil de jugadores
grandes con tiro exterior.*

*# En la NBA moderna, la posición de pivot está atravesando una transformación similar.
Cada vez se valoran más los pívots capaces de amenazar ofensivamente desde
el exterior, abriendo así el campo, aumentando los espacios y obligando a las
defensas a salir de la pintura. Este cambio permite que las ofensivas sean
más dinámicas y versátiles.*

*# # Por ello, a continuación, analizaremos cómo esta primera componente
principal, que refleja la capacidad de los jugadores para anotar eficazmente
desde todas las zonas del campo, influye en el salario de los pívots en la NBA actual.*

#Voy a probar un modelo de regresion lineal con la posicion pivot

```
lmpivot<-lm(Salario~Dim.1 , data =nba2_filtrado[nba2_filtrado$POS=="C",] )  
summary(lmpivot)
```

```

# Filtramos solo los "C"
centers <- nba2_filtrado[nba2_filtrado$POS == "C", ]

ggplot(centers, aes(x = Dim.1, y = Salario)) +
  geom_point(color = "#2c7fb8", alpha = 0.7, size = 2) +
  geom_smooth(method = "lm", se = TRUE, color = "darkred") +
  theme_minimal() +
  labs(
    title = "Relación entre Dim.1 y Salario (Posición: Center)",
    x = "Dim.1",
    y = "Salario"
  )

#Voy a probar un árbol con las dos componentes principales y la eficiencia

# Árbol de regresión, el valor cp = 0.004135472 sale de tomar cp = 0.000100000
# y mediante cptable "podar" a un árbol más sencillo sin perder mucha precisión
arbol_salario <- rpart(Salario ~ Dim.1+Dim.2+eficiencia, data = centers,
                      method = "anova", control = rpart.control(cp = 0.004135472))

#Al árbol se le añade la variable eficiencia, porque como ya hemos visto
#en otras cosas es una gran variable para clasificar de manera inicial
#a jugadores realmente buenos de los q no los on
# Visualizar el árbol
rpart.plot(arbol_salario,
           type = 3,
           fallen.leaves = TRUE,
           box.palette = "Blues",
           shadow.col = "gray",
           main = "Árbol de regresión")

#trabajamos con un cp pequeño y podemos con la información de la tabla

arbol_salario$cptable

# Predicciones
centers$pred_arbol <- predict(arbol_salario)

# Error cuadrático medio (RMSE)
rmse <- sqrt(mean((centers$Salario - centers$pred_arbol)^2))

```

```

print(paste("RMSE:", round(rmse, 2)))

ggplot(centers, aes(x = Salario, y = pred_arbol)) +
  geom_point(color = "#1f78b4", alpha = 0.7, size = 2) + # puntos más visibles
  geom_abline(slope = 1, intercept = 0, linetype = "dashed", color = "darkred", size = 1)
  ↪ + # línea ideal
labs(
  title = "Predicción del Salario con Árbol de Regresión",
  subtitle = "Jugadores con posición Center (C)",
  x = "Salario real",
  y = "Salario predicho por el árbol"
) +
theme_minimal(base_size = 14) +
theme(
  plot.title = element_text(face = "bold", hjust = 0.5),
  plot.subtitle = element_text(hjust = 0.5)
)

```