

Álvaro del Cerro Hernández



*Who is the Real King of NBA*

Proyecto EDA - DataScience

Contenido

[INTRODUCCIÓN 2](#_Toc70450107)

[OBTENCIÓN DE DATOS 2](#_Toc70450108)

[HIPOTESIS 3](#_Toc70450109)

[EXPLORACION 4](#_Toc70450110)

[CONCLUSION 6](#_Toc70450111)

# INTRODUCCIÓN

NBA, National Basketball Association es la liga privada de baloncesto profesional que se disputa en Estados Unidos. A lo largo de su historia, muchos jugadores han destacado pero los más reconocidos por todo el mundo han sido Michael Jordan, Kobe Bryant y Lebron James. Estos tres jugadores han roto todo tipo de records, alcanzando números extraordinarios y destacando por encima de cualquier otro jugador profesional. Son numerosas las comparaciones entre los tres, destacando fortalezas y debilidades que impliquen diferencias para situar a un jugador por encima del otro. Diferentes épocas, diferentes modos de juego y estilos son los factores por los que existe la gran pregunta: ¿Quién es el mejor jugador de la historia?

Para intentar dar solución a esta pregunta, vamos a realizar una exploración de los datos registrados de estos tres jugadores, vamos a compararlos y establecer criterios para analizar la posibilidad de establecer a un claro vencedor. Se establecerán criterios acordes al mundo de baloncesto, temporadas jugadas, títulos obtenidos, números totales, desempeño de la temporada y desempeño contra equipos de la liga.

# OBTENCIÓN DE DATOS

Necesitamos encontrar datos y estadísticas de los números obtenidos de los 3 jugadores a lo largo de su carrera profesional en la NBA.

Para llevar a cabo el proyecto vamos a emplear 4 bases de datos obtenido del siguiente link : <https://www.kaggle.com/xvivancos/michael-jordan-kobe-bryant-and-lebron-james-stats>

Las bases de datos contienen todas las estadísticas de los tres jugadores. Para cada base de datos nos centraremos en estadísticas concretas para analizar nuestras hipótesis y poder comparar entre los tres jugadores. Muchos de los datos se repiten entre las bases debido a que las variables son las mismas, un ejemplo es la agrupación de puntos, rebotes, asistencias por temporada. Esto se tendrá en cuenta para evitar duplicados.

Las bases de datos son las siguientes: “allgames\_stats”, “salaries”, game\_high\_stats”, y “per\_game\_stats”. Dado que en las bases de datos las variables se repiten, recogeremos variables únicas de cada base para trabajar las hipótesis descritas.

Las principales bases de datos y la descripción de principales variables únicas por base empleadas son:

* “allgames\_stats” y “per\_game\_stats”: todos los datos de todos los partidos de los tres jugadores
  + X: partido en casa o fuera
  + Result: resultado
  + Opp: oponente G : games
  + MP: minutes played
  + FG: field goals
  + FGA: field goal attempts
  + 3P: 3 point field goals
  + 2P: 2 point field goals
  + FT: free throws
  + ORB: offensive rebounds
  + DRB: defensive rebounds
  + PF: personal fouls
  + PTS: points
* “game\_highs\_stats”: dato del mejor partido jugado en cada temporada
  + GmSc: game score
* “salaries”: datos sobre los ingresos totales que han obtenido cada jugador en cada temporada disputada.

# HIPOTESIS

La principal hipótesis es lograr saber quién ha sido el mejor jugador de la historia de la NBA. Para ello vamos a comparar los datos básicos, total puntos, total rebotes, total asistencias etc. Para lograr comparar adecuadamente a los tres jugadores y obtener valores cuantitativos sobre los datos, vamos a crear un sistema de puntos. Este sistema trata de dar un valor a cada variable por jugador, es decir, cuando comparemos los datos de una variable establecida, daremos un valor a cada jugador. Al final y para establecer la conclusión, sumaremos todos los resultados obtenidos y veremos si conseguimos identificar a un ganador claro.

El sistema de puntos trata de lo siguiente: se asignará un valor de 3, 2 o 1 según el resultado de la variable, siendo 3 el jugador que tenga mejor número y 1 el que tenga el número más bajo.

Cabe destacar que hay variables que son más cualitativas que cuantitativas, por motivos de percepción, pensamiento o interpretación. Todas estas variables serán resaltadas y justificadas pero no contarán como parte de la hipótesis principal

Hipótesis secundarias que aportaran criterio a la hipótesis principal son:

* Quien es el jugador más ofensivo
* Quien es el jugador más defensivo
* Quien es el jugador que juega más en equipo
* Quien tiene mayor porcentaje de aciertos
* Quien es el jugador que más minutos ha jugado
* Quien es el jugador que más partidos ha jugado

Con los datos obtenidos podemos también plantearnos las siguientes preguntas para resolver la principal hipótesis en otros ámbitos.

Para el videojuego oficial 2K NBA, según los datos obtenidos, podemos llegar a desarrollar un perfil más ajustado para las características de los tres jugadores. Se pueden destacar aquellas variables y obtener una calificación global para ajustar más el perfil digital del jugador acorde al juego.

Hay que destacar que dos de estos tres jugadores, Kobe Bryant y Lebron James han coincidido en su carrera profesional, mientras que Michael Jordan es un jugador que marcó un antes y un después en la historia del baloncesto.

Las épocas de estos jugadores son distintas, esto implica variables cualitativas que no servirán como criterio para la hipótesis dado su complejidad. Variables como el material deportivo, la equipación y vestimenta, el material del campo, conocimiento sobre el desarrollo muscular entre otros, son factores que caracterizan épocas y se entiende que hay cierta diferencia en el ámbito deportivo que favorece a las nuevas generaciones. No obstante, estos factores no serán evaluados y no aportarán criterio objetivo más que simple conocimiento.

# EXPLORACION

Comenzamos con el análisis de datos y exploración.

Importamos las librerías necesarias para el análisis siendo numpy, pandas, matplotlib.pyplot y seaborn. Para todos los notebooks queremos conocer las características de las bases de datos, por lo que imprimiremos shape, info, columns, dtypes, describe y campos nulos.

En primer lugar exploramos la base de datos “per\_game\_stats”. Esta base de datos contiene agrupado por temporada jugada de cada jugador, la media de las cifras que han obtenido, un ejemplo sería los puntos que lebron james ha obtenido de media en la temporada 2010-2011. Cada columna representa una variable obtenida del jugador. Estas variables incluyen puntos, minutos jugados, rebotes, etc.

Para esta exploración queremos identificar los valores de cada jugador según las siguientes columnas: minutos jugados (MP), canastas metidas (FG), canastas de 3P y su porcentaje de aciertos (3P y 3P%), canastas de 2P y su porcentaje de aciertos (2P y 2P%), tapones (BLK), puntos (PTS), asistencias (AST), robos (STL), rebotes ofensivos, defensivos y totales (ORB, DRB, TRB) Para esto creamos una función donde seleccionamos la columna que queremos, hacemos un groupby de la columna jugador y obtenemos las estadísticas descriptivas. Repetimos para cada columna y cogemos la media para evaluar nuestra hipótesis.

Para tener una representación más visual y analizar cualitativamente, creamos 4 gráficos distplot. Cada gráfico identifica la columna que queremos mostrar e identifica a los jugadores. Los 4 gráficos son puntos, asistencias, rebotes y canastas.

En segundo lugar, exploraremos la base de datos “game\_highs\_stats”, base que contiene por cada temporada jugada de cada jugador, el registro más alto de cada columna, es decir, para la temporada 2010-2011 de lebron james, los puntos mas altos obtenidos en un partido, el mayor numero de asistencias dadas en un partido, etc.

Para esta exploración vamos a realizar los mismos pasos que en la primera exploración de la base de datos “per\_game\_stats” con la simple diferencia que vamos a separar los valores por la columna “RSorPO” (temporada regular o playoffs). Esta separación la realizamos ya que el rendimiento de los jugadores es diferente tanto por el oponente como por la motivación del partido. Esto también se puede analizar cualitativamente.

Una vez segmentado por la columna “RSorPO”, exploramos las columnas canastas (FG), triples (3P), dobles (2P) y puntos (PTS). La diferencia respecto a la exploración anterior es que esta vez nos fijaremos en el estadístico max.

En tercer lugar, vamos a explorar la base de datos “allgames\_stats”, que contiene todos los registros de los partidos donde han participado. Aquí queremos destacar y observar las victorias y derrotas respecto a los oponentes, además de los registros comunes (puntos, asistencias...)

Para la exploración los mismos pasos que en las exploraciones previas. Esta vez exploraremos toda la base por jugador, y por tanto utilizamos la función loc donde la columna “jugador” es igual al jugador que queremos analizar. Una vez tenemos solo los datos de un jugador, exploramos los estadísticos min, max y media de las columnas PTS, AST, TRB y FG.

A continuación creamos un boxplot de todos los partidos donde utilizamos las columnas Date y PTS. Así podemos observar el rendimiento en cuanto a puntos del jugador. También creamos un gráfico donde observamos la comparación entre victorias y derrotas por oponente. Este gráfico conlleva un análisis cualitativo donde podemos respaldar las conclusiones con datos.

Repetimos los mismos gráficos pero esta vez segmentamos los datos del jugador por la columna “X” (juega en casa o fuera) para ver el rendimiento del jugador. Total, obtenemos 4 graficos, dos con representación del rendimiento jugando como local, y 2 jugando como visitante.

Repetimos todo lo anterior pero con los otros dos jugadores restantes.

Por último, como exploración adicional por curiosidad, exploramos la base de datos “salaries”. Queremos ver la comparativa de sueldos entre los tres. Para ello, una vez visto las características de la base, tenemos que transformar la columna salario de dtype object a dtype int, para poder sumarlo. También tenemos que eliminar el “$”.

Para ello creamos una columna nueva, realizamos los cambios previos y la concatenamos con la base principal. Una vez realizado todo, mostramos el salario máximo, mínimo y la suma total de cada jugador.

Adicional creamos un catplot que indica cuantos salarios tenemos registrados, es decir, temporadas con registro; y otro catplot para ver la distribución de los salarios de cada jugador.

Esta exploración no es mas que curiosidad y no aporta valor a la conclusión.

# CONCLUSION

Vamos a basar la conclusión de nuestra hipótesis principal en dos, conclusión cuantitativa y conclusión cualitativa. Para la conclusión cuantitativa vamos a emplear todos los datos obtenidos y valorados con el sistema de puntos creados. Para la conclusión cualitativa, analizaremos todos los gráficos obtenidos y valoraremos un criterio homogéneo y parcial, pero no afecta a nuestra hipótesis dado el factor interpretación

En cuanto a la conclusión cuantitativa, observamos lo siguiente:



En la izquierda podemos ver los valores explorados y en la tabla de la derecha nuestro sistema de puntos.

Después de nuestro análisis podemos concluir que para nuestra hipótesis principal, la corona se la lleva Michael Jordan.

Para obtener conclusión de las hipótesis secundarias, tenemos que observar la tabla y relacionarla con la hipótesis. Por ejemplo, el jugador con más minutos de media es Lebron, el jugador con más asistencias es Lebron, el jugador más ofensivo (mayor media de puntos) es Jordan.