Universidad ORT de Montevideo

Facultad de Ingeniería

Big Data Obligatorio

Agustina Goñi– 284421

Docente: Eduardo García

2024

Contenido

Contenido

1. Caso de estudio

En el presente documento se analiza un conjunto de datos tabulares relacionados con la NBA, que contiene información sobre jugadores, equipos, partidos y estadísticas acumuladas. Este conjunto de datos permite explorar múltiples aspectos del baloncesto profesional, como el rendimiento de jugadores y equipos, características físicas y estadísticas individuales, así como resultados y tendencias de los partidos.

Los datos utilizados abarcan diversas temporadas de la NBA y fueron proporcionados por el docente como parte de la actividad académica. Estos registros representan una valiosa fuente para realizar análisis exploratorios y visualizaciones, con el objetivo de responder preguntas clave sobre el desempeño y la evolución de la liga.

2. Planteamiento del problema

El análisis de los datos de la NBA tiene como objetivo responder preguntas clave relacionadas con el rendimiento de jugadores, equipos y partidos. Estas preguntas permitirán explorar patrones y tendencias significativos en el conjunto de datos, proporcionando información relevante para entender mejor la dinámica de la liga.

A continuación, se presentan las 8 preguntas seleccionadas:

- ¿Cuáles son los jugadores nacidos fuera de Estados Unidos que llegaron al All-Star (jugando al menos un minuto) o al Hall of Fame?
- ¿Cuáles son las posiciones del Draft que tienen más jugadores en el Hall of Fame?
- ¿Cuáles son las 10 universidades que han producido la mayor cantidad de jugadores en el Hall of Fame?
- ¿Cuáles son los 10 equipos con las mejores temporadas en términos de número de victorias?
- ¿Quiénes son los entrenadores con más premios, y cómo se distribuyen los premios entre ellos?
- ¿Cuáles fueron los movimientos de los 10 entrenadores más galardonados entre equipos a lo largo de su carrera?
- ¿Cómo han evolucionado la altura y el peso promedio de los jugadores a lo largo de las décadas?
- ¿Cómo ha cambiado la relevancia de cada posición en el Draft a lo largo de las décadas?

3. Análisis exploratorio vía pandas

En este análisis exploratorio, utilice Pandas y Spark para comprender la estructura y el contenido de los datos relacionados con la NBA. Ambas herramientas ofrecen métodos eficientes para manipular y explorar datos, y esta comparación permite resaltar las diferencias en su uso y resultados.

3.1 Carga de datos

El primer paso del análisis fue cargar los datos desde archivos CSV con información sobre la NBA. Además se, manejaron errores y se estandarizaron los nombres de las columnas convirtiéndolos a minúsculas y eliminando los espacios en blanco. Esto facilita trabajar con los datos durante el análisis.

Con Pandas
Los datos fueron cargados localmente desde archivos CSV utilizando pandas.read_csv(). Este enfoque permite cargar los datos directamente en memoria, ideal para conjuntos pequeños o medianos. Se utilizó Python para procesar cada archivo, renombrar columnas y asegurarse de que los datos estuvieran listos para el análisis.

Con Spark
En Spark, los datos fueron cargados en un entorno distribuido utilizando spark.read.csv() después de haberlos transferido al sistema de archivos HDFS mediante scp. Este método es ideal para manejar grandes volúmenes de datos, ya que permite procesarlos de manera distribuida y escalable.
Cada archivo fue leído como un DataFrame y almacenado en un diccionario para su posterior uso.

3.2 Dimensiones y columnas

En esta etapa, se realizó un análisis básico para identificar las dimensiones y las columnas presentes en cada tabla del conjunto de datos. Esto permitió comprender la estructura y la información contenida en los datos.

En los resultados, podremos observar las dimensiones de las tablas analizadas, lo que nos permite conocer cuántos registros (filas) y atributos (columnas) posee cada una. Por ejemplo, si una tabla tiene una dimensión de (40.3), esto significa que contiene 40 registros y 3 atributos.

Además, las columnas representan los atributos o variables que describen cada registro en la tabla. Los nombres de estas columnas nos brindan una idea del tipo de datos que alberga cada tabla.

Pandas

```
Tabla: abbrev
Dimensiones: (40, 3)
Columnas: ['abbrev_type', 'code', 'full_name']
================================================
Tabla: awards_coaches
Dimensiones: (61, 5)
Columnas: ['year', 'coachID', 'award', 'lgID', 'note']
================================================
Tabla: awards_players
Dimensiones: (1719, 6)
Columnas: ['playerID', 'award', 'year', 'lgID', 'note', 'pos']
================================================
Tabla: coaches
Dimensiones: (1689, 9)
Columnas: ['coachID', 'year', 'tmID', 'lgID', 'stint', 'won', 'lost', 'post_wins', 'post_losses']
================================================
Tabla: draft
Dimensiones: (9003, 11)
Columnas: ['draftYear', 'draftRound', 'draftSelection', 'draftOverall', 'tmID', 'firstName', 'lastName', 'suffixName', 'playerID', 'draftFrom', 'lgID']
================================================
Tabla: hof
Dimensiones: (328, 4)
Columnas: ['year', 'hofID', 'name', 'category']
================================================
Tabla: master
Dimensiones: (5061, 26)
Columnas: ['bioID', 'useFirst', 'firstName', 'middleName', 'lastName', 'nameGiven', 'fullGivenName', 'nameSuffix', 'nameNick', 'pos', 'firstseason', 'la
stseason', 'height', 'weight', 'college', 'collegeOther', 'birthDate', 'birthCity', 'birthState', 'birthCountry', 'highSchool', 'hsCity', 'hsState', 'hs
Country', 'deathDate', 'race']
================================================
Tabla: player_allstar
Dimensiones: (1609, 23)
Columnas: ['player_id', 'last_name', 'first_name', 'season_id', 'conference', 'league_id', 'games_played', 'minutes', 'points', 'o_rebounds', 'd_rebound
s', 'rebounds', 'assists', 'steals', 'blocks', 'turnovers', 'personal_fouls', 'fg_attempted', 'fg_made', 'ft_attempted', 'ft_made', 'three_attempted',
'three_made']
================================================
Tabla: series_post
Dimensiones: (775, 9)
Columnas: ['year', 'round', 'series', 'tmIDWinner', 'lgIDWinner', 'tmIDLoser', 'lgIDLoser', 'W', 'L']
================================================
Tabla: teams
Dimensiones: (1536, 60)
Columnas: ['year', 'lgID', 'tmID', 'franchID', 'confID', 'divID', 'rank', 'confRank', 'playoff', 'name', 'o_fgm', 'o_fga', 'o_ftm', 'o_fta', 'o_3pm', 'o
_3pa', 'o_oreb', 'o_dreb', 'o_reb', 'o_asts', 'o_pf', 'o_stl', 'o_to', 'o_blk', 'o_pts', 'd_fgm', 'd_fga', 'd_ftm', 'd_fta', 'd_3pm', 'd_3pa', 'd_oreb',
'd_dreb', 'd_reb', 'd_asts', 'd_pf', 'd_stl', 'd_to', 'd_blk', 'd_pts', 'o_tmRebound', 'd_tmRebound', 'homeWon', 'homeLost', 'awayWon', 'awayLost', 'neu
tWon', 'neutLoss', 'confWon', 'confLoss', 'divWon', 'divLoss', 'pace', 'won', 'lost', 'games', 'min', 'arena', 'attendance', 'bbtmID']
================================================
```

Spark

```
Tabla: basketball_abbrev
Dimensiones: (40, 3)
Columnas: ['abbrev_type', 'code', 'full_name']
================================================
Tabla: basketball_awards_coaches
Dimensiones: (61, 5)
Columnas: ['year', 'coachID', 'award', 'lgID', 'note']
================================================
Tabla: basketball_awards_players
Dimensiones: (1719, 6)
Columnas: ['playerID', 'award', 'year', 'lgID', 'note', 'pos']
================================================
Tabla: basketball_coaches
Dimensiones: (1689, 9)
Columnas: ['coachID', 'year', 'tmID', 'lgID', 'stint', 'won', 'lost', 'post_wins', 'post_losses']
================================================
Tabla: basketball_draft
Dimensiones: (9003, 11)
Columnas: ['draftYear', 'draftRound', 'draftSelection', 'draftOverall', 'tmID', 'firstName', 'lastName', 'suffixName', 'playerID', 'draftFrom', 'lgID']
================================================
Tabla: basketball_hof
Dimensiones: (328, 4)
Columnas: ['year', 'hofID', 'name', 'category']
================================================
Tabla: basketball_master
Dimensiones: (5061, 26)
Columnas: ['bioID', 'useFirst', 'firstName', 'middleName', 'lastName', 'nameGiven', 'fullGivenName', 'nameSuffix', 'nameNick', 'pos', 'firstseason', 'las
tseason', 'height', 'weight', 'college', 'collegeOther', 'birthDate', 'birthCity', 'birthState', 'birthCountry', 'highSchool', 'hsCity', 'hsState', 'hsCo
untry', 'deathDate', 'race']
================================================
Tabla: basketball_player_allstar
Dimensiones: (1609, 23)
Columnas: ['player_id', 'last_name', 'first_name', 'season_id', 'conference', 'league_id', 'games_played', 'minutes', 'points', 'o_rebounds', 'd_rebound
s', 'rebounds', 'assists', 'steals', 'blocks', 'turnovers', 'personal_fouls', 'fg_attempted', 'fg_made', 'ft_attempted', 'ft_made', 'three_attempted', 't
hree_made']
================================================
Tabla: basketball_series_post
Dimensiones: (775, 9)
Columnas: ['year', 'round', 'series', 'tmIDWinner', 'lgIDWinner', 'tmIDLoser', 'lgIDLoser', 'W', 'L']
================================================
Tabla: basketball_teams
Dimensiones: (1536, 60)
Columnas: ['year', 'lgID', 'tmID', 'franchID', 'confID', 'divID', 'rank', 'confRank', 'playoff', 'name', 'o_fgm', 'o_fga', 'o_ftm', 'o_fta', 'o_3pm', 'o_
3pa', 'o_oreb', 'o_dreb', 'o_reb', 'o_asts', 'o_pf', 'o_stl', 'o_to', 'o_blk', 'o_pts', 'd_fgm', 'd_fga', 'd_ftm', 'd_fta', 'd_3pm', 'd_3pa', 'd_oreb',
'd_dreb', 'd_reb', 'd_asts', 'd_pf', 'd_stl', 'd_to', 'd_blk', 'd_pts', 'o_tmRebound', 'd_tmRebound', 'homeWon', 'homeLost', 'awayWon', 'awayLost', 'neut
Won', 'neutLoss', 'confWon', 'confLoss', 'divWon', 'divLoss', 'pace', 'won', 'lost', 'games', 'min', 'arena', 'attendance', 'bbtmID']
================================================
```

Tabla abbrev:

- **Dimensiones:** (40 filas, 3 columnas).
- **Descripción:** Contiene información sobre abreviaturas utilizadas en otras tablas.
- **Columnas:**
  - abbrev_type: El tipo de abreviatura (puede ser equipo, liga, etc.).
  - code: El código abreviado.
  - full_name: El nombre completo asociado al código.

Tabla awards_coaches:

- **Dimensiones:** (61 filas, 5 columnas).
- **Descripción:** Lista de premios otorgados a entrenadores a lo largo de los años.
- **Columnas:**
  - year: El año en que se otorgó el premio.
  - coachID: Identificador único del entrenador.
  - award: Nombre del premio.
  - lgID: Liga a la que pertenece el entrenador.
  - note: Notas adicionales.

Tabla awards_players:

- **Dimensiones:** (1719 filas, 6 columnas).
- **Descripción:** Premios ganados a jugadores.
- **Columnas:**
  - playerID: Identificador único del jugador.
  - award: Nombre del premio.
  - year: Año en que se otorgó el premio.
  - lgID: Liga del jugador.
  - note: Notas adicionales.
  - pos: Posición del jugador.

Tabla coaches:

- **Dimensiones:** (1689 filas, 9 columnas).
- **Descripción:** Estadísticas de entrenadores por temporada.
- **Columnas:**
  - coachID: Identificador único del entrenador.
  - year: Año de la temporada.
  - tmID: Identificador del equipo.
  - lgID: Liga del entrenador.
  - stint: Número de veces que el entrenador lideró el equipo esa temporada.
  - won: Partidos ganados en la temporada regular.
  - lost: Partidos perdidos en la temporada regular.
  - post_wins: Partidos ganados en los playoffs.
  - post_losses: Partidos perdidos en los playoffs.

Tabla draft:

- **Dimensiones:** (9003 filas, 11 columnas).
- **Descripción:** Información sobre los jugadores seleccionados en el draft.
- **Columnas:**
  - draftYear: Año del draft.
  - draftRound: Ronda del draft.

- draftNumber: Número de elección en la ronda.
- teamID: Equipo que seleccionó al jugador.
- playerID: Identificador único del jugador.
- college: Universidad del jugador.
- notes: Notas adicionales sobre el jugador o el draft.
- lgID: Liga del equipo seleccionador.
- pickOverall: Número total de selección general.
- pickTeamID: Identificador del equipo seleccionador.
- teamNotes: Notas relacionadas con el equipo.

Tabla hof (Hall of Fame):

- **Dimensiones:** (328 filas, 4 columnas).
- **Descripción:** Lista de miembros destacados en el Hall of Fame (Salón de la fama).
- **Columnas:**
  - year: Año de inclusión.
  - hofID: Identificador único del miembro.
  - name: Nombre del miembro.
  - category: Categoría (ej. jugador, entrenador).

Tabla master:

- **Dimensiones:** (5061 filas, 26 columnas).
- **Descripción:** Información detallada sobre los jugadores.
  - playerID: Identificador único del jugador.
  - firstName: Nombre del jugador.
  - lastName: Apellido del jugador.
  - givenName: Nombre completo del jugador.
  - birthDate: Fecha de nacimiento.
  - deathDate: Fecha de fallecimiento (si aplica).
  - college: Universidad a la que asistió el jugador.
  - height: Altura del jugador (en pulgadas).
  - weight: Peso del jugador (en libras).
  - birthCity: Ciudad de nacimiento.
  - birthState: Estado de nacimiento.
  - birthCountry: País de nacimiento.
  - deathCity: Ciudad de fallecimiento (si aplica).
  - deathState: Estado de fallecimiento (si aplica).
  - deathCountry: País de fallecimiento (si aplica).
  - highSchool: Escuela secundaria del jugador.
  - position: Posición en la que jugaba.
  - shoots: Lado de disparo (derecho o izquierdo).
  - draftTeamID: Equipo que seleccionó al jugador en el draft.
  - draftYear: Año del draft.
  - draftRound: Ronda del draft.
  - draftPick: Elección del draft.
  - debut: Fecha del debut en la NBA.
  - finalGame: Fecha del último partido en la NBA.
  - careerLength: Duración de la carrera en años.
  - hof: Indicador de si está en el Hall of Fame.

Tabla player_allstar:

- **Dimensiones:** (1609 filas, 23 columnas).
- **Descripción:** Estadísticas de jugadores en el All-Star Game.
- **Columnas:**
  - player_id: Identificador único del jugador.
  - last_name: Apellido del jugador.
  - first_name: Nombre del jugador.
  - season_id: Identificador de la temporada.
  - conference: Conferencia en la que participó el jugador.
  - league_id: Liga en la que jugó el jugador.
  - games_played: Número de partidos jugados en el All-Star.
  - minutes: Minutos jugados en el All-Star.
  - points: Puntos anotados en el All-Star.
  - o_rebounds: Rebotes ofensivos en el All-Star.
  - d_rebounds: Rebotes defensivos en el All-Star.
  - rebounds: Rebotes totales en el All-Star.
  - assists: Asistencias realizadas en el All-Star.
  - steals: Robos de balón en el All-Star.
  - blocks: Bloqueos realizados en el All-Star.
  - turnovers: Pérdidas de balón en el All-Star.
  - personal_fouls: Faltas personales en el All-Star.
  - fg_attempted: Tiros de campo intentados en el All-Star.
  - fg_made: Tiros de campo anotados en el All-Star.
  - ft_attempted: Tiros libres intentados en el All-Star.
  - ft_made: Tiros libres anotados en el All-Star.
  - three_attempted: Tiros de tres puntos intentados en el All-Star.
  - three_made: Tiros de tres puntos anotados en el All-Star.

Tabla series_post:

- **Dimensiones:** (775 filas, 9 columnas).
- **Descripción:** Detalles de series de playoffs.
- **Columnas:**
  - tmIDWinner / tmIDLoser: Equipos ganador y perdedor.
  - W / L: Partidos ganados y perdidos en la serie.

Tabla teams:

- **Dimensiones:** (1536 filas, 60 columnas).
- **Descripción:** Estadísticas detalladas de equipos por temporada.
- **Columnas:**
  - won / lost: Partidos ganados y perdidos.
  - attendance: Asistencia total.

3.3 Tipo de datos

Se verificaron los tipos de datos presentes en cada tabla. Esto es importante para entender el formato de las variables y asegurarse de que los datos sean adecuados para el análisis.

```
Tabla: abbrev
Tipos de datos:
abbrev_type    object
code           object
full_name      object
dtype: object
==================================================
Tabla: awards_coaches
Tipos de datos:
year        int64
coachID    object
award      object
lgID       object
note       object
dtype: object
==================================================
Tabla: awards_players
Tipos de datos:
```

```
Tabla: basketball_abbrev
Tipos de datos:
abbrev_type: StringType()
code: StringType()
full_name: StringType()
==================================================
Tabla: basketball_awards_coaches
Tipos de datos:
year: IntegerType()
coachID: StringType()
award: StringType()
lgID: StringType()
note: StringType()
--------------------------------------------------
```

En Pandas, se observa que en la tabla abbrev todas las columnas (abbrev_type, code, full_name) tienen el tipo de dato object, lo que indica que contienen texto o valores categóricos. En Spark, estas mismas columnas son de tipo StringType(), indicando también que almacenan texto.

En la tabla awards_coaches, Pandas clasifica la columna year como numérica (int64), mientras que las demás (coachID, award, etc.) son categóricas (object). En Spark, year es de tipo IntegerType() y las demás columnas son de tipo StringType(). Ambas herramientas coinciden en que la tabla contiene una combinación de datos numéricos y categóricos.

```
==================================================
Tabla: coaches
Tipos de datos:
coachID         object
year             int64
tmID            object
lgID            object
stint            int64
won            float64
lost           float64
post_wins      float64
post_losses    float64
dtype: object
==================================================
 ==========================================
 Tabla: basketball_coaches
 Tipos de datos:
 coachID: StringType()
 year: IntegerType()
 tmID: StringType()
 lgID: StringType()
 stint: IntegerType()
 won: IntegerType()
 lost: IntegerType()
 post_wins: IntegerType()
 post_losses: IntegerType()
 ==========================================
```

En Pandas, la tabla coaches combina datos numéricos como won y lost (int64) con datos categóricos como coachID y tmID (object). En Spark, los datos numéricos son de tipo IntegerType() y los categóricos, como coachID y tmID, son de tipo StringType(). Ambas herramientas reflejan una combinación de datos numéricos y categóricos relacionados con el rendimiento de los entrenadores.

```
dtype: object
==================================================
Tabla: draft
Tipos de datos:
draftYear          int64
draftRound         int64
draftSelection     int64
draftOverall       int64
tmID              object
firstName         object
lastName          object
suffixName        object
playerID          object
draftFrom         object
lgID              object
dtype: object
==================================================
```

En Pandas, en la tabla draft predominan los datos categóricos como firstName, lastName y tmID (object), junto con datos numéricos como draftYear y draftOverall (int64), lo que refleja información relacionada con el proceso de selección de jugadores. En Spark, los datos numéricos como draftYear y draftOverall son de tipo IntegerType(), mientras que los

categóricos como firstName, lastName y tmID son de tipo StringType(). Ambos enfoques coinciden en su combinación de datos categóricos y numéricos.

```
Tabla: hof
Tipos de datos:
year         int64
hofID       object
name        object
category    object
dtype: object
---------------------------------
Tabla: basketball_hof
Tipos de datos:
year: IntegerType()
hofID: StringType()
name: StringType()
category: StringType()
```

En la tabla hof, se encuentran columnas como year, que es de tipo int64 en Pandas y IntegerType() en Spark, y columnas categóricas como name y category, que son de tipo object en Pandas y StringType() en Spark. Esto refleja una combinación de datos numéricos y categóricos que describen a las personas incluidas en el Salón de la Fama.

```
-------------------------------------
Tabla: master
Tipos de datos:
bioID             object
useFirst          object
firstName         object
middleName        object
lastName          object
nameGiven         object
fullGivenName     object
nameSuffix        object
nameNick          object
pos               object
firstseason      float64
lastseason       float64
height           float64
weight           float64
college           object
collegeOther      object
```

```
==================================================
Tabla: basketball_master
Tipos de datos:
bioID: StringType()
useFirst: StringType()
firstName: StringType()
middleName: StringType()
lastName: StringType()
nameGiven: StringType()
fullGivenName: StringType()
nameSuffix: StringType()
nameNick: StringType()
pos: StringType()
firstseason: IntegerType()
lastseason: IntegerType()
height: DoubleType()
weight: IntegerType()
college: StringType()
collegeOther: StringType()
birthDate: StringType()
birthCity: StringType()
birthState: StringType()
birthCountry: StringType()
highSchool: StringType()
hsCity: StringType()
hsState: StringType()
hsCountry: StringType()
deathDate: StringType()
race: StringType()
```

En la tabla master, la mayoría de las columnas son categóricas, como bioID, firstName y birthCity, que son de tipo object en Pandas y StringType() en Spark. Además, columnas como firstseason y lastseason son numéricas, clasificadas como float64 en Pandas y IntegerType() en Spark, mientras que height y weight también son numéricas, siendo float64 en Pandas y DoubleType() y IntegerType() respectivamente en Spark. Esta tabla combina datos categóricos y numéricos, lo que es común en conjuntos con información detallada sobre personas.

```
----------------------------------------
Tabla: player_allstar
Tipos de datos:
player_id           object
last_name           object
first_name          object
season_id            int64
conference          object
league_id           object
games_played         int64
minutes              int64
points             float64
o_rebounds         float64
d_rebounds         float64
rebounds           float64
assists            float64
steals             float64
blocks             float64
```

```
==================================================
Tabla: basketball_player_allstar
Tipos de datos:
player_id: StringType()
last_name: StringType()
first_name: StringType()
season_id: IntegerType()
conference: StringType()
league_id: StringType()
games_played: IntegerType()
minutes: IntegerType()
points: IntegerType()
o_rebounds: IntegerType()
d_rebounds: IntegerType()
rebounds: IntegerType()
assists: IntegerType()
steals: IntegerType()
blocks: IntegerType()
turnovers: IntegerType()
personal_fouls: IntegerType()
fg_attempted: IntegerType()
fg_made: IntegerType()
ft_attempted: IntegerType()
ft_made: IntegerType()
three_attempted: IntegerType()
three_made: IntegerType()
--------------------------------------------------
```

En la tabla player_allstar, se observan tipos de datos variados. Algunas columnas, como player_id y conference, son de tipo object en Pandas y StringType() en Spark, mientras que las relacionadas con el rendimiento de los jugadores, como games_played y minutes, son de tipo int64 en Pandas e IntegerType() en Spark. Esta tabla incluye también datos numéricos como points y rebounds, que son de tipo int64 en pandas e IntegerType() en spark, reflejando la diversidad de los datos estadísticos de los jugadores.

```
Tabla: series_post
Tipos de datos:
year           int64
round         object
series        object
tmIDWinner    object
lgIDWinner    object
tmIDLoser     object
lgIDLoser     object
W              int64
L              int64
dtype: object
==================================================
Tabla: basketball_series_post
Tipos de datos:
year: IntegerType()
round: StringType()
series: StringType()
tmIDWinner: StringType()
lgIDWinner: StringType()
tmIDLoser: StringType()
lgIDLoser: StringType()
W: IntegerType()
L: IntegerType()
==================================================
```

En la tabla series_post, se identifica una combinación de datos categóricos y numéricos. Las columnas year, W y L son de tipo int64 en Pandas e IntegerType() en Spark, representando datos numéricos. Por otro lado, columnas como round, series y los identificadores de equipos y ligas (tmIDWinner, lgIDWinner, etc.) son de tipo object en Pandas y StringType() en Spark, lo que indica que contienen datos categóricos. Esta estructura es adecuada para representar información tanto textual como numérica sobre las series de playoffs.

```
awayWon          int64
awayLost         int64
neutWon          int64
neutLoss         int64
confWon          int64
confLoss         int64
divWon           int64
divLoss          int64
pace             int64
won              int64
lost             int64
games            int64
min            float64
arena           object
attendance       int64
bbtmID          object
dtype: object
```

```
==================================================
Tabla: basketball_teams
Tipos de datos:
year: IntegerType()
lgID: StringType()
tmID: StringType()
franchID: StringType()
confID: StringType()
divID: StringType()
rank: IntegerType()
confRank: IntegerType()
playoff: StringType()
name: StringType()
o_fgm: IntegerType()
o_fga: IntegerType()
o_ftm: IntegerType()
o_fta: IntegerType()
o_3pm: IntegerType()
o_3pa: IntegerType()
o_oreb: IntegerType()
o_dreb: IntegerType()
o_reb: IntegerType()
o_asts: IntegerType()
o_pf: IntegerType()
o_stl: IntegerType()
o_to: IntegerType()
o_blk: IntegerType()
o_pts: IntegerType()
d_fgm: IntegerType()
d_fga: IntegerType()
d_ftm: IntegerType()
d_fta: IntegerType()
d_3pm: IntegerType()
d_3pa: IntegerType()
d_oreb: IntegerType()
d_dreb: IntegerType()
d_reb: IntegerType()
d_asts: IntegerType()
d_pf: IntegerType()
d_stl: IntegerType()
d_to: IntegerType()
d_blk: IntegerType()
d_pts: IntegerType()
o_tmRebound: IntegerType()
d_tmRebound: IntegerType()
```

En la tabla teams, predomina el uso de datos numéricos, como points, rebounds y assists, que son de tipo int64 en Pandas e IntegerType() en Spark, para estadísticas del equipo. Las columnas categóricas como tmID y name son de tipo object en Pandas y StringType() en Spark, proporcionando información para identificar a los equipos y describir su desempeño y características.

3.4 Valores nulos

Para identificar la calidad y completitud de los datos, se analizaron los valores nulos presentes en cada columna de las tablas.

A continuación, se muestra y describe el resultado:

```
Tabla: abbrev
Valores nulos por columna:
abbrev_type    0
code           0
full_name      0
dtype: int64
==================================================
Tabla: awards_coaches
Valores nulos por columna:
year        0
coachID     0
award       0
lgID        0
note       57
dtype: int64
```

```
Tabla: basketball_abbrev
Valores nulos por columna:
+-----------+----+---------+
|abbrev_type|code|full_name|
+-----------+----+---------+
|          0|   0|        0|
+-----------+----+---------+

Tabla: basketball_awards_coaches
Valores nulos por columna:
+----+-------+-----+----+----+
|year|coachID|award|lgID|note|
+----+-------+-----+----+----+
|   0|      0|    0|   0|  57|
+----+-------+-----+----+----+
```

En la tabla abbrev, no se detectaron valores nulos en ninguna de las columnas, lo que garantiza que los datos en esta tabla están completamente disponibles y no requieren tratamiento adicional.

En la tabla awards_coaches, se observó que la columna note contiene 57 valores nulos, mientras que las demás columnas no presentan valores faltantes. Esto indica que las notas adicionales son opcionales y podrían no estar presentes en todos los registros.

```
==================================================
Tabla: awards_players
Valores nulos por columna:
playerID       0
award          0
year           0
lgID           0
note        1682
pos          886
dtype: int64
==================================================

Tabla: basketball_awards_players
Valores nulos por columna:
+--------+-----+----+----+----+---+
|playerID|award|year|lgID|note|pos|
+--------+-----+----+----+----+---+
|       0|    0|   0|   0|1682|886|
+--------+-----+----+----+----+---+
```

La tabla awards_players presenta valores nulos en las columnas note (1682) y pos (886). Esto sugiere que las notas y las posiciones no están disponibles para todos los jugadores, lo que podría requerir una imputación o exclusión dependiendo del análisis.

```
==============================================
Tabla: coaches
Valores nulos por columna:
coachID         0
year            0
tmID            0
lgID            0
stint           0
won             9
lost            9
post_wins      40
post_losses    40
dtype: int64


  Tabla: basketball_coaches
  Valores nulos por columna:
  +-------+----+----+----+-----+---+----+---------+-----------+
  |coachID|year|tmID|lgID|stint|won|lost|post_wins|post_losses|
  +-------+----+----+----+-----+---+----+---------+-----------+
  |      0|   0|   0|   0|    0|  9|   9|       40|         40|
  +-------+----+----+----+-----+---+----+---------+-----------+
```

En la tabla coaches, las columnas relacionadas con victorias (won, lost) tienen 9 valores nulos, mientras que las columnas de rendimiento en postemporada (post_wins, post_losses) presentan 40 valores nulos. Esto indica que algunos entrenadores no tienen datos completos de rendimiento, posiblemente porque no participaron en postemporada.

```
==========================================
Tabla: draft
Valores nulos por columna:
draftYear             0
draftRound            0
draftSelection        0
draftOverall          0
tmID                  0
firstName             0
lastName              0
suffixName         9001
playerID           5189
draftFrom            25
lgID                  0
dtype: int64
```

```
Tabla: basketball_draft
Valores nulos por columna:
+---------+----------+--------------+------------+----+---------+--------+----------+--------+---------+----+
|draftYear|draftRound|draftSelection|draftOverall|tmID|firstName|lastName|suffixName|playerID|draftFrom|lgID|
+---------+----------+--------------+------------+----+---------+--------+----------+--------+---------+----+
|        0|         0|             0|           0|   0|        0|       0|      9001|    5189|        6|   0|
+---------+----------+--------------+------------+----+---------+--------+----------+--------+---------+----+
```

En la tabla draft, se identificaron valores nulos significativos en columnas como suffixName (9001), playerID (5189) y draftFrom (25). La alta cantidad de valores faltantes en suffixName sugiere que no es un dato crítico para la mayoría de los registros.

```
==================================
Tabla: hof
Valores nulos por columna:
year          0
hofID       188
name          0
category      0
dtype: int64
==================================
  Tabla: basketball_hof
  Valores nulos por columna:
  +----+-----+----+--------+
  |year|hofID|name|category|
  +----+-----+----+--------+
  |   0|  188|   0|       0|
  +----+-----+----+--------+
```

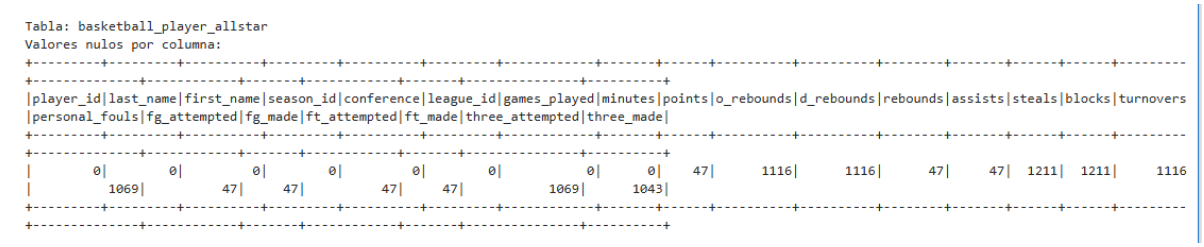
En la tabla hof, la columna hofID presenta 188 valores nulos, lo que podría indicar registros sin un identificador único para el Salón de la Fama.

```
==========================================
Tabla: master
Valores nulos por columna:
bioID              0
useFirst         831
firstName         24
middleName      1816
lastName           0
nameGiven       5052
fullGivenName   5035
nameSuffix      4738
nameNick        2354
pos              182
firstseason       15
lastseason        15
height            13
weight            14
college          178
collegeOther    4208
birthDate         11
birthCity       1174
birthState      1454
birthCountry    1160
highSchool       773
hsCity           765
hsState          834
hsCountry        765
deathDate          0
race             159
dtype: int64
```

```
Tabla: basketball_master
Valores nulos por columna:
+-----+--------+---------+----------+--------+---------+------------+----------+--------+---+----------+----------+------+------+-------+-----------+-
--------+---------+----------+-----------+---------+-----+-------+---------+---------+----+
|bioID|useFirst|firstName|middleName|lastName|nameGiven|fullGivenName|nameSuffix|nameNick|pos|firstseason|lastseason|height|weight|college|collegeOther|b
irthDate|birthCity|birthState|birthCountry|highSchool|hsCity|hsState|hsCountry|deathDate|race|
+-----+--------+---------+----------+--------+---------+------------+----------+--------+---+----------+----------+------+------+-------+-----------+-
--------+---------+----------+-----------+---------+-----+-------+---------+---------+----+
|    0|     831|       24|      1816|       0|     5052|        5035|      4738|    2354|182|        15|        15|    13|    14|    178|       4208|
11|     1174|      1454|        1160|       773|   765|    834|      765|        0| 159|
+-----+--------+---------+----------+--------+---------+------------+----------+--------+---+----------+----------+------+------+-------+-----------+-
--------+---------+----------+-----------+---------+-----+-------+---------+---------+----+
```

La tabla master tiene múltiples columnas con valores nulos. Algunas de las más relevantes son middleName (1816), nameNick (2354) y collegeOther (4208). También se identificaron valores nulos en datos clave como firstseason (15) y birthCity (1174), lo que podría complicar ciertos análisis.

```
------------------------------------------
Tabla: player_allstar
Valores nulos por columna:
player_id             0
last_name             0
first_name            0
season_id             0
conference            0
league_id             0
games_played          0
minutes               0
points               47
o_rebounds         1116
d_rebounds         1116
rebounds             47
assists              47
steals             1211
blocks             1211
turnovers          1116
personal_fouls     1069
fg_attempted         47
fg_made              47
ft_attempted         47
ft_made              47
three_attempted    1069
three_made         1043
dtype: int64
==============================
```



```
Tabla: basketball_player_allstar
Valores nulos por columna:
+---------+---------+----------+---------+----------+---------+------------+-------+------+----------+----------+--------+-------+------+------+---------
+-------------+------------+-----+----------+-----+-------------+---------+
|player_id|last_name|first_name|season_id|conference|league_id|games_played|minutes|points|o_rebounds|d_rebounds|rebounds|assists|steals|blocks|turnovers
|personal_fouls|fg_attempted|fg_made|ft_attempted|ft_made|three_attempted|three_made|
+---------+---------+----------+---------+----------+---------+------------+-------+------+----------+----------+--------+-------+------+------+---------
+-------------+------------+-----+----------+-----+-------------+---------+
|       0|       0|         0|        0|         0|        0|           0|      0|    47|      1116|      1116|      47|     47|  1211|  1211|     1116
|         1069|         47|     47|         47|     47|         1069|     1043|
+---------+---------+----------+---------+----------+---------+------------+-------+------+----------+----------+--------+-------+------+------+---------
+-------------+------------+-----+----------+-----+-------------+---------+
```

En la tabla player_allstar, se detectaron valores nulos en columnas relacionadas con estadísticas del jugador, como points (47), rebounds (47) y steals (1211). Esto indica que no todos los jugadores tienen estadísticas completas en los partidos All-Star.

```
==========================================
Tabla: series_post
Valores nulos por columna:
year          0
round         0
series        0
tmIDWinner    0
lgIDWinner    0
tmIDLoser     6
lgIDLoser     0
W             0
L             0
dtype: int64
```

```
Tabla: basketball_series_post
Valores nulos por columna:
+----+-----+------+----------+----------+---------+---------+---+---+
|year|round|series|tmIDWinner|lgIDWinner|tmIDLoser|lgIDLoser|  W|  L|
+----+-----+------+----------+----------+---------+---------+---+---+
|   0|    0|     0|         0|         0|        6|        0|  0|  0|
+----+-----+------+----------+----------+---------+---------+---+---+
```

En la tabla series_post, la única columna con valores nulos es tmIDLoser (6), lo que podría significar que no se registró un equipo perdedor en ciertos casos.

```
Tabla: teams
Valores nulos por columna:
year            0
lgID            0
tmID            0
franchID        0
confID        472
divID          38
rank            0
confRank        0
playoff       635
name            0
o_fgm           0
o_fga           0
o_ftm           0
o_fta           0
o_3pm           0
o_3pa           0
awayWon         0
awayLost        0
neutWon         0
neutLoss        0
confWon         0
confLoss        0
divWon          0
divLoss         0
pace            0
won             0
lost            0
games           0
min           214
arena         189
attendance      0
bbtmID        220
dtype: int64
```

```
Tabla: basketball_teams
Valores nulos por columna:
```

year	lgID	tmID	franchID	confID	divID	rank	confRank	playoff	name	o_fgm	o_fga	o_ftm	o_fta	o_3pm	o_3pa	o_oreb	o_dreb	o_reb	o_asts
0	0	0	0	472	38	0	0	635	0	0	0	0	0	0	0	0	0	0	0

o_pf	o_stl	o_to	o_blk	o_pts	d_fgm	d_fga	d_ftm	d_fta	d_3pm	d_3pa	d_oreb	d_dreb	d_reb	d_asts	d_pf	d_stl	d_to	d_blk	d_pts
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

o_tmRebound	d_tmRebound	homeWon	homeLost	awayWon	awayLost	neutWon	neutLoss	confWon	confLoss	divWon	divLoss	pace	won	lost	games	min	arena	attendance	bbtmID
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	214	189	0	220

Finalmente, en la tabla teams, las columnas confID (472), divID (38) y playoff (635) presentan valores nulos, lo que podría reflejar equipos sin información de conferencias o desempeño en playoffs. Además, columnas como min (214) y arena (189) también tienen valores faltantes.

3.5 Vista previa de datos

Para comprender mejor la estructura y el contenido de los datos en cada tabla, se generó una vista previa de las primeras filas.

En esta etapa recorro cada DataFrame, imprimo su nombre y muestro las primeras cinco filas de sus datos. Esto permite observar cómo están estructuradas las tablas y los valores iniciales de cada columna. A continuación están los resultados.

```
Tabla: abbrev
Vista previa de las primeras filas:
  abbrev_type code           full_name
0       Round   DF     Division Finals
1  Conference   EC  Eastern Conference
2  Conference   WC  Western Conference
3    Division   AT   Atlantic Division
4    Division   ED    Eastern Division
==================================================
```

```
Vista previa de basketball_abbrev:
+-----------+----+------------------+
|abbrev_type|code|         full_name|
+-----------+----+------------------+
|      Round|  DF|   Division Finals|
| Conference|  EC|Eastern Conference|
| Conference|  WC|Western Conference|
|   Division|  AT| Atlantic Division|
|   Division|  ED|  Eastern Division|
+-----------+----+------------------+
only showing top 5 rows
```

```
==================================================
Tabla: awards_coaches
Vista previa de las primeras filas:
   year    coachID                 award lgID note
0  1962  gallaha01  NBA Coach of the Year  NBA  NaN
1  1963  hannual01  NBA Coach of the Year  NBA  NaN
2  1964  auerbre01  NBA Coach of the Year  NBA  NaN
3  1965  schaydo01  NBA Coach of the Year  NBA  NaN
4  1966   kerrjo01  NBA Coach of the Year  NBA  NaN
==================================================


Vista previa de basketball_awards_coaches:
+----+---------+--------------------+----+----+
|year|  coachID|               award|lgID|note|
+----+---------+--------------------+----+----+
|1962|gallaha01|NBA Coach of the ...| NBA|null|
|1963|hannual01|NBA Coach of the ...| NBA|null|
|1964|auerbre01|NBA Coach of the ...| NBA|null|
|1965|schaydo01|NBA Coach of the ...| NBA|null|
|1966| kerrjo01|NBA Coach of the ...| NBA|null|
+----+---------+--------------------+----+----+
only showing top 5 rows
==================================================
Tabla: awards_players
Vista previa de las primeras filas:
    playerID               award  year lgID note  pos
0  feeribo01  All-NBA First Team  1946  NBA  NaN  NaN
1  fulksjo01  All-NBA First Team  1946  NBA  NaN  NaN
2  mckinho01  All-NBA First Team  1946  NBA  NaN  NaN
3  miasest01  All-NBA First Team  1946  NBA  NaN  NaN
4  zasloma01  All-NBA First Team  1946  NBA  NaN  NaN
==================================================
Vista previa de basketball_awards_players:
+---------+------------------+----+----+----+----+
| playerID|             award|year|lgID|note| pos|
+---------+------------------+----+----+----+----+
|feeribo01|All-NBA First Team|1946| NBA|null|null|
|fulksjo01|All-NBA First Team|1946| NBA|null|null|
|mckinho01|All-NBA First Team|1946| NBA|null|null|
|miasest01|All-NBA First Team|1946| NBA|null|null|
|zasloma01|All-NBA First Team|1946| NBA|null|null|
+---------+------------------+----+----+----+----+
only showing top 5 rows
```

```
==================================================
Tabla: coaches
Vista previa de las primeras filas:
     coachID  year tmID  lgID  stint   won  lost  post_wins  post_losses
0  johnsne01  1961  PGR  ABL1      1  41.0  40.0        0.0          1.0
1  auerbre01  1946  WSC   NBA      1  49.0  11.0        2.0          4.0
2  birchpa01  1946  PIT   NBA      1  15.0  45.0        0.0          0.0
3  cliffro01  1946  CLR   NBA      2  13.0  10.0        1.0          2.0
4  cohalne01  1946  NYK   NBA      1  33.0  27.0        2.0          3.0
==================================================

  Vista previa de basketball_coaches:
  +---------+----+----+----+-----+---+----+---------+-----------+
  |  coachID|year|tmID|lgID|stint|won|lost|post_wins|post_losses|
  +---------+----+----+----+-----+---+----+---------+-----------+
  |johnsne01|1961| PGR|ABL1|    1| 41|  40|        0|          1|
  |auerbre01|1946| WSC| NBA|    1| 49|  11|        2|          4|
  |birchpa01|1946| PIT| NBA|    1| 15|  45|        0|          0|
  |cliffro01|1946| CLR| NBA|    2| 13|  10|        1|          2|
  |cohalne01|1946| NYK| NBA|    1| 33|  27|        2|          3|
  +---------+----+----+----+-----+---+----+---------+-----------+
  only showing top 5 rows


  Tabla: draft
  Vista previa de las primeras filas:
     draftYear  draftRound  draftSelection  draftOverall tmID firstName  \
  0       1967           0               0             0  ANA   Darrell
  1       1967           0               0             0  ANA       Bob
  2       1967           0               0             0  ANA       Bob
  3       1967           0               0             0  ANA      Mike
  4       1967           0               0             0  ANA       Tom

    lastName suffixName   playerID       draftFrom lgID
  0    Hardy        NaN  hardyda01          Baylor  ABA
  1  Krulish        NaN        NaN         Pacific  ABA
  2    Lewis        NaN  lewisbo01  North Carolina  ABA
  3     Lynn        NaN   lynnmi01            UCLA  ABA
  4  Workman        NaN  workmto01         Seattle  ABA
  ==================================================

Vista previa de basketball_draft:
+---------+----------+-------------+-----------+----+---------+--------+----------+---------+--------------+----+
|draftYear|draftRound|draftSelection|draftOverall|tmID|firstName|lastName|suffixName| playerID|     draftFrom|lgID|
+---------+----------+-------------+-----------+----+---------+--------+----------+---------+--------------+----+
|     1967|         0|            0|          0| ANA|  Darrell|   Hardy|      null|hardyda01|        Baylor| ABA|
|     1967|         0|            0|          0| ANA|      Bob| Krulish|      null|     null|       Pacific| ABA|
|     1967|         0|            0|          0| ANA|      Bob|   Lewis|      null|lewisbo01|North Carolina| ABA|
|     1967|         0|            0|          0| ANA|     Mike|    Lynn|      null| lynnmi01|          UCLA| ABA|
|     1967|         0|            0|          0| ANA|      Tom| Workman|      null|workmto01|       Seattle| ABA|
+---------+----------+-------------+-----------+----+---------+--------+----------+---------+--------------+----+
only showing top 5 rows
```

```
==================================================
Tabla: hof
Vista previa de las primeras filas:
   year      hofID               name     category
0  1959        NaN  Amos Alonzo Stagg  Contributor
1  1959        NaN      Charles Hyatt       Player
2  1959        NaN      Edward Hickox  Contributor
3  1959  mikange01       George Mikan       Player
4  1959        NaN      Hank Luisetti       Player
==================================================

Vista previa de basketball_hof:
+----+---------+----------------+-----------+
|year|    hofID|            name|   category|
+----+---------+----------------+-----------+
|1959|     null|Amos Alonzo Stagg|Contributor|
|1959|     null|    Charles Hyatt|     Player|
|1959|     null|    Edward Hickox|Contributor|
|1959|mikange01|     George Mikan|     Player|
|1959|     null|    Hank Luisetti|     Player|
+----+---------+----------------+-----------+
only showing top 5 rows
```

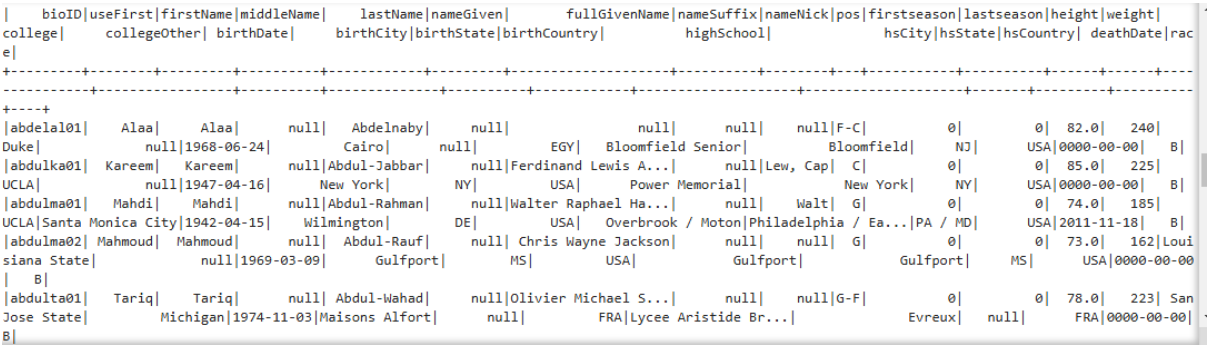
```
==================================================
Tabla: master
Vista previa de las primeras filas:
       bioID useFirst firstName middleName      lastName nameGiven  \
0  abdelal01     Alaa      Alaa        NaN     Abdelnaby       NaN
1  abdulka01   Kareem    Kareem        NaN  Abdul-Jabbar       NaN
2  abdulma01    Mahdi     Mahdi        NaN  Abdul-Rahman       NaN
3  abdulma02  Mahmoud   Mahmoud        NaN    Abdul-Rauf       NaN
4  abdulta01    Tariq     Tariq        NaN   Abdul-Wahad       NaN

                  fullGivenName nameSuffix  nameNick  pos  ...   birthDate  \
0                           NaN        NaN       NaN  F-C  ...  1968-06-24
1  Ferdinand Lewis Alcindor, Jr.        NaN  Lew, Cap    C  ...  1947-04-16
2    Walter Raphael Hazzard, Jr.        NaN      Walt    G  ...  1942-04-15
3            Chris Wayne Jackson        NaN       NaN    G  ...  1969-03-09
4     Olivier Michael Saint-Jean        NaN       NaN  G-F  ...  1974-11-03
```

	birthCity	birthState	birthCountry	highSchool	\
0	Cairo	NaN	EGY	Bloomfield Senior	
1	New York	NY	USA	Power Memorial	
2	Wilmington	DE	USA	Overbrook / Moton	
3	Gulfport	MS	USA	Gulfport	
4	Maisons Alfort	NaN	FRA	Lycee Aristide Briand	

	hsCity	hsState	hsCountry	deathDate	race
0	Bloomfield	NJ	USA	0000-00-00	B
1	New York	NY	USA	0000-00-00	B
2	Philadelphia / Easton	PA / MD	USA	2011-11-18	B
3	Gulfport	MS	USA	0000-00-00	B
4	Evreux	NaN	FRA	0000-00-00	B

[5 rows x 26 columns]



bioID	useFirst	firstName	middleName	lastName	nameGiven	fullGivenName	nameSuffix	nameNick	pos	firstseason	lastseason	height	weight	college	collegeOther	birthDate	birthCity	birthState	birthCountry	highSchool	hsCity	hsState	hsCountry	deathDate	race
abdelal01	Alaa	Alaa	null	Abdelnaby	null	null	null	null	F-C	0	0	82.0	240	Duke	null	1968-06-24	Cairo	null	EGY	Bloomfield Senior	Bloomfield	NJ	USA	0000-00-00	B
abdulka01	Kareem	Kareem	null	Abdul-Jabbar	null	Ferdinand Lewis A...	null	Lew, Cap	C	0	0	85.0	225	UCLA	null	1947-04-16	New York	NY	USA	Power Memorial	New York	NY	USA	0000-00-00	B
abdulma01	Mahdi	Mahdi	null	Abdul-Rahman	null	Walter Raphael Ha...	null	Walt	G	0	0	74.0	185	UCLA	Santa Monica City	1942-04-15	Wilmington	DE	USA	Overbrook / Moton	Philadelphia / Ea...	PA / MD	USA	2011-11-18	B
abdulma02	Mahmoud	Mahmoud	null	Abdul-Rauf	null	Chris Wayne Jackson	null	null	G	0	0	73.0	162	Louisiana State	null	1969-03-09	Gulfport	MS	USA	Gulfport	Gulfport	MS	USA	0000-00-00	B
abdulta01	Tariq	Tariq	null	Abdul-Wahad	null	Olivier Michael S...	null	null	G-F	0	0	78.0	223	San Jose State	Michigan	1974-11-03	Maisons Alfort	null	FRA	Lycee Aristide Br...	Evreux	null	FRA	0000-00-00	B

Tabla: player_allstar

Vista previa de las primeras filas:

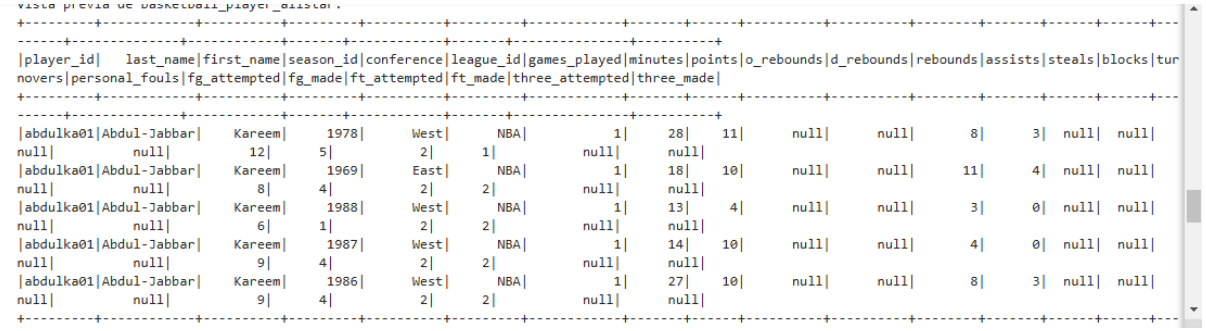
	player_id	last_name	first_name	season_id	conference	league_id	\
0	abdulka01	Abdul-Jabbar	Kareem	1978	West	NBA	
1	abdulka01	Abdul-Jabbar	Kareem	1969	East	NBA	
2	abdulka01	Abdul-Jabbar	Kareem	1988	West	NBA	
3	abdulka01	Abdul-Jabbar	Kareem	1987	West	NBA	
4	abdulka01	Abdul-Jabbar	Kareem	1986	West	NBA	

	games_played	minutes	points	o_rebounds	...	steals	blocks	turnovers	\
0	1	28	11.0	NaN	...	NaN	NaN	NaN	
1	1	18	10.0	NaN	...	NaN	NaN	NaN	
2	1	13	4.0	NaN	...	NaN	NaN	NaN	
3	1	14	10.0	NaN	...	NaN	NaN	NaN	
4	1	27	10.0	NaN	...	NaN	NaN	NaN	

```
   personal_fouls  fg_attempted  fg_made  ft_attempted  ft_made  \
0             NaN          12.0      5.0           2.0      1.0
1             NaN           8.0      4.0           2.0      2.0
2             NaN           6.0      1.0           2.0      2.0
3             NaN           9.0      4.0           2.0      2.0
4             NaN           9.0      4.0           2.0      2.0

   three_attempted  three_made
0              NaN         NaN
1              NaN         NaN
2              NaN         NaN
3              NaN         NaN
4              NaN         NaN

[5 rows x 23 columns]
```



```
|player_id|   last_name|first_name|season_id|conference|league_id|games_played|minutes|points|o_rebounds|d_rebounds|rebounds|assists|steals|blocks|turnovers|personal_fouls|fg_attempted|fg_made|ft_attempted|ft_made|three_attempted|three_made|
|abdulka01|Abdul-Jabbar|    Kareem|     1978|      West|      NBA|           1|     28|    11|      null|      null|       8|      3|  null|  null|     null|          null|          12|      5|           2|      1|           null|      null|
|abdulka01|Abdul-Jabbar|    Kareem|     1969|      East|      NBA|           1|     18|    10|      null|      null|      11|      4|  null|  null|     null|          null|           8|      4|           2|      2|           null|      null|
|abdulka01|Abdul-Jabbar|    Kareem|     1988|      West|      NBA|           1|     13|     4|      null|      null|       3|      0|  null|  null|     null|          null|           6|      1|           2|      2|           null|      null|
|abdulka01|Abdul-Jabbar|    Kareem|     1987|      West|      NBA|           1|     14|    10|      null|      null|       4|      0|  null|  null|     null|          null|           9|      4|           2|      2|           null|      null|
|abdulka01|Abdul-Jabbar|    Kareem|     1986|      West|      NBA|           1|     27|    10|      null|      null|       8|      3|  null|  null|     null|          null|           9|      4|           2|      2|           null|      null|
```

```
==================================================
Tabla: series_post
Vista previa de las primeras filas:
   year round series tmIDWinner lgIDWinner tmIDLoser lgIDLoser  W  L
0  1946     F      O        PHW        NBA       CHS       NBA  4  1
1  1946    QF      M        NYK        NBA       CLR       NBA  2  1
2  1946    QF      M        PHW        NBA       STB       NBA  2  1
3  1946    SF      N        PHW        NBA       NYK       NBA  2  0
4  1946    SF      N        CHS        NBA       WSC       NBA  4  2
==================================================

Vista previa de basketball_series_post:
+----+-----+------+----------+----------+---------+---------+---+---+
|year|round|series|tmIDWinner|lgIDWinner|tmIDLoser|lgIDLoser|  W|  L|
+----+-----+------+----------+----------+---------+---------+---+---+
|1946|    F|     O|       PHW|       NBA|      CHS|      NBA|  4|  1|
|1946|   QF|     M|       NYK|       NBA|      CLR|      NBA|  2|  1|
|1946|   QF|     M|       PHW|       NBA|      STB|      NBA|  2|  1|
|1946|   SF|     N|       PHW|       NBA|      NYK|      NBA|  2|  0|
|1946|   SF|     N|       CHS|       NBA|      WSC|      NBA|  4|  2|
+----+-----+------+----------+----------+---------+---------+---+---+
only showing top 5 rows
```

```
==================================================
Tabla: teams
Vista previa de las primeras filas:
   year lgID tmID franchID confID divID  rank  confRank playoff  \
0  1946  NBA  BOS      BOS    NaN    ED     5         0     NaN
1  1946  NBA  CHS      CHS    NaN    WD     1         0       F
2  1946  NBA  CLR      CLR    NaN    WD     3         0      R1
3  1946  NBA  DTF      DTF    NaN    WD     4         0     NaN
4  1946  NBA  NYK      NYK    NaN    ED     3         0      SF

                name  ...  divWon  divLoss  pace  won  lost  games      min  \
0    Boston Celtics  ...      11       19     0   22    38     60  14500.0
1     Chicago Stags  ...      17        8     0   39    22     61  14840.0
2  Cleveland Rebels  ...      10       14     0   30    30     60  14600.0
3   Detroit Falcons  ...       8       16     0   20    40     60  14600.0
4   New York Knicks  ...      13       17     0   33    27     60  14575.0

                         arena  attendance  bbtmID
0                Boston Garden       32767     BOS
1              Chicago Stadium           0     CHS
2              Cleveland Arena           0     CLR
3              Detroit Olympia           0     DTF
4  Madison Square Garden (III)       32767     NYK

[5 rows x 60 columns]
```

```
Tabla: basketball_teams
Valores nulos por columna:
+----+----+----+--------+------+-----+----+--------+-------+----+-----+-----+-----+-----+-----+-----+------+------+-----+------+----+-----+----+-----+---
--+-----+-----+-----+-----+-----+-----+------+------+-----+-----+----+-----+----+-----+-----+----------+----------+------+--------+-------+--------+-
-----+--------+-------+--------+-----+-----+----+---+----+-----+---+-----+----------+------+
|year|lgID|tmID|franchID|confID|divID|rank|confRank|playoff|name|o_fgm|o_fga|o_fta|o_3pm|o_3pa|o_oreb|o_dreb|o_reb|o_asts|o_pf|o_stl|o_to|o_blk|o_p
ts|d_fgm|d_fga|d_ftm|d_fta|d_3pm|d_3pa|d_oreb|d_dreb|d_reb|d_asts|d_pf|d_stl|d_to|d_blk|d_pts|o_tmRebound|d_tmRebound|homeWon|homeLost|awayWon|awayLost|n
eutWon|neutLoss|confWon|confLoss|divWon|divLoss|pace|won|lost|games|min|arena|attendance|bbtmID|
+----+----+----+--------+------+-----+----+--------+-------+----+-----+-----+-----+-----+-----+-----+------+------+-----+------+----+-----+----+-----+---
--+-----+-----+-----+-----+-----+-----+------+------+-----+-----+----+-----+----+-----+-----+----------+----------+------+--------+-------+--------+-
-----+--------+-------+--------+-----+-----+----+---+----+-----+---+-----+----------+------+
|   0|   0|   0|      0|   472|   38|   0|      0|   635|   0|   0|   0|   0|   0|   0|   0|    0|    0|   0|    0|   0|   0|   0|   0|   0|   0|
0|   0|   0|   0|   0|   0|   0|   0|   0|   0|   0|   0|   0|   0|   0|   0|        0|        0|     0|      0|     0|      0|     0|
0|      0|      0|      0|      0|      0|   0|  0|   0|   0|214|  189|        0|   220|
+----+----+----+--------+------+-----+----+--------+-------+----+-----+-----+-----+-----+-----+-----+------+------+-----+------+----+-----+----+-----+---
--+-----+-----+-----+-----+-----+-----+------+------+-----+-----+----+-----+----+-----+-----+----------+----------+------+--------+-------+--------+-
-----+--------+-------+--------+-----+-----+----+---+----+-----+---+-----+----------+------+
```

3.6 MER


abbrev
code varchar
abbrev_type varchar
full_name varchar
awards_coaches
coachID varchar
year int
award varchar
lgID varchar
note varchar
awards_players
playerID varchar
award varchar
year int
lgID varchar
note varchar
pos varchar
coaches
coachID varchar
year int
tmID varchar
lgID varchar
stint int
won int
lost int
post_wins int
post_losses int
draft
playerID varchar
draftYear int
draftRound int
draftSelection int
draftOverall int
tmID varchar
firstName varchar
lastName varchar
suffixName varchar
draftFrom varchar
lgID varchar
hof
hofID varchar
name varchar
year int
category varchar
master
bioID varchar
useFirst varchar
firstName varchar
middleName varchar
lastName varchar
nameGiven varchar
fullGivenName varchar
nameSuffix varchar
nameNick varchar
pos varchar
firstseason int
lastseason int
height int
weight int
college varchar
collegeOther varchar
birthDate date
birthCity varchar
birthState varchar
birthCountry varchar
highSchool varchar
hsCity varchar
hsState varchar
hsCountry varchar
deathDate date
race varchar
player_allstar
player_id varchar
last_name varchar
first_name varchar
season_id int
conference varchar
league_id varchar
games_played int
minutes int
points int
o_rebounds int
d_rebounds int
rebounds int
assists int
steals int
blocks int
turnovers int
personal_fouls int
fg_attempted int
fg_made int
ft_attempted int
ft_made int
three_attempted int
three_made int
teams
tmID varchar
year int
lgID varchar
franchID varchar
confID varchar
divID varchar
rank int
confRank int
playoff varchar
name varchar
o_pts int
d_pts int
series_post
year int
round varchar
series varchar
tmIDWinner varchar
lgIDWinner varchar
tmIDLoser varchar
lgIDLoser varchar
W int
L int

3.7 Limpieza de tablas

- La tabla master se limpió para asegurarse de que sea útil y clara para responder las preguntas del análisis. Por ejemplo, la columna birthcountry es muy importante para la Pregunta 1, ya que ayuda a identificar a los jugadores nacidos fuera de Estados Unidos. Durante la limpieza, se usó un mapeo completo para corregir y uniformar los valores de esta columna. Esto incluyó arreglos manuales de códigos incorrectos, como convertir ENG a United Kingdom, y agregar una lista completa de códigos de países oficiales (ISO 3166-1) para completar los datos. Si algún código no pudo ser reconocido, se reemplazó con "Desconocido". Así, los datos sobre los países de nacimiento quedaron claros y consistentes.
  Otras columnas también fueron revisadas para responder preguntas clave. Por ejemplo, la columna college, que se usa en la Pregunta 3 para analizar qué universidades aportaron más jugadores al Hall of Fame, fue ajustada para que los valores vacíos se llenaran con "Desconocido". La columna birthdate, que es esencial para la Pregunta 6, permitió calcular las décadas en las que nacieron los jugadores y descubrir tendencias en sus características físicas, como altura y peso.
  Las columnas numéricas height y weight también fueron importantes para este análisis. Se corrigieron valores inválidos, como alturas o pesos poco realistas, y se convirtieron las unidades a un formato estándar (centímetros para altura y kilogramos para peso). Además, se eliminaron valores extremos que no tenían sentido, asegurando que los datos fueran confiables y representaran la realidad de los jugadores. En cuanto a la columna bioID, que es clave para la Pregunta 8, se revisó para asegurarse de que no tuviera duplicados ni datos vacíos. Esto fue importante porque esta columna sirve como identificador principal para conectar los datos con otras tablas. También se trabajó en la columna pos, normalizando las combinaciones de posiciones. Por ejemplo, combinaciones como C-F y F-C se unificaron para que tuvieran un solo formato, facilitando el análisis de la evolución de las posiciones en el Draft.
- La tabla hof fue limpiada para garantizar que los datos sean relevantes y consistentes con el análisis necesario. Se seleccionaron las columnas hofid, year y category, ya que permiten identificar a los jugadores incluidos en el Hall of Fame y el año de su inclusión. La columna hofid es esencial porque sirve como clave primaria para cruzar datos con otras tablas, como la de jugadores, mientras que year proporciona una dimensión temporal para el análisis. Inicialmente, se incluyó category para filtrar únicamente a los jugadores, ya que otras categorías, como entrenadores, no son relevantes para el análisis. En el proceso de limpieza, se normalizó la columna category eliminando espacios y convirtiendo los valores a minúsculas para asegurar la consistencia. Se filtraron exclusivamente los registros correspondientes a jugadores, descartando datos de otras categorías.
  Posteriormente, se eliminaron los registros con valores nulos en hofid, ya que esta columna es la clave primaria y debía garantizar su unicidad. Además, el formato de hofid fue estandarizado eliminando espacios y convirtiendo los valores a minúsculas, lo que facilita su uso en cruces de datos. Tras asegurar que no había duplicados en hofid, la columna category fue eliminada, ya que ya no era necesaria para el análisis.

Con estas acciones, la tabla hof quedó optimizada para ser utilizada en el análisis relacionado con los jugadores destacados en el Hall of Fame, asegurando datos limpios y consistentes.

- La tabla draft fue limpiada para garantizar que los datos sean consistentes y relevantes para el análisis. Se seleccionaron las columnas playerID, draftOverall y draftYear, ya que son esenciales para responder preguntas específicas. La columna playerID es necesaria para cruzar los datos del draft con los jugadores en la Pregunta 2, donde se analizan las posiciones del draft con más jugadores en el Hall of Fame. La columna draftOverall se utiliza para identificar las posiciones específicas en el draft, también en la Pregunta 2. Finalmente, la columna draftYear permite calcular las décadas del draft, lo cual es importante para la Pregunta 8, que analiza la evolución de la importancia de las posiciones a lo largo del tiempo. Durante la limpieza, se eliminaron los registros con valores nulos en playerID, ya que esta columna es clave para los cruces. Además, se normalizó el formato de playerID, eliminando espacios y convirtiendo los valores a minúsculas para asegurar consistencia. En la columna draftOverall, se convirtieron los valores a numéricos, descartando aquellos que no eran válidos o eran menores o iguales a 0. También se eliminaron duplicados en las combinaciones de playerID, draftYear y draftOverall para evitar redundancia en el análisis.

- La tabla teams fue limpiada para asegurar la consistencia y calidad de los datos, dado su uso en preguntas específicas del análisis. Se seleccionaron las columnas tmid, won, lost, games, year y name, ya que son fundamentales para responder a las preguntas planteadas. La columna tmid se utiliza para identificar de manera única a los equipos, lo cual es necesario en las Preguntas 4 y 7. Las columnas won y lost permiten calcular el número de victorias y derrotas, esenciales para la Pregunta 4, donde se analiza la relación entre victorias y derrotas de los equipos más exitosos. Por su parte, la columna name proporciona los nombres completos de los equipos, importantes para la visualización en la Pregunta 7, que estudia los movimientos de entrenadores entre equipos. Durante la limpieza, se normalizó el formato de tmid, eliminando espacios y convirtiendo los valores a minúsculas para garantizar la consistencia. Se eliminaron registros con valores nulos en tmid y name, ya que son claves para el análisis. También se verificaron y eliminaron duplicados en las combinaciones de tmid y year, evitando redundancias. En las columnas won y lost, se eliminaron valores negativos, y se verificó la coherencia de la columna games, asegurando que coincidiera con la suma de won y lost. Los registros inconsistentes fueron identificados y descartados, asegurando datos confiables.

- La tabla player_allstar fue limpiada con el objetivo de responder a la Pregunta 1, que analiza qué jugadores extranjeros llegaron a participar en el All-Star. Se seleccionaron las columnas player_id, first_name, last_name, season_id y minutes porque son esenciales para identificar a los jugadores, mostrar sus nombres, detallar las temporadas en las que participaron y considerar su tiempo de juego.
  La columna player_id se usó como clave única para identificar a los jugadores y cruzar esta tabla con otras, como master, para obtener información adicional sobre su país de origen. Las columnas first_name y last_name permiten mostrar los nombres completos de los jugadores en el análisis. La columna season_id se incluyó

para identificar las temporadas específicas en las que participaron en el All-Star, mientras que minutes sirvió como criterio para filtrar a los jugadores que participaron activamente, asegurando que hayan jugado al menos un minuto.
Durante la limpieza, se normalizó el formato de player_id eliminando espacios y convirtiendo los valores a minúsculas, lo que garantiza la consistencia en los datos. Se eliminaron los registros con valores nulos en las columnas player_id, season_id y minutes, ya que son datos imprescindibles para el análisis. También se filtraron los jugadores que hayan jugado al menos un minuto, asegurando que solo se incluyan participantes activos. Después de esto, se eliminaron duplicados para dejar un único registro por jugador y se descartó la columna minutes tras su uso en el filtro, ya que no era necesaria para el análisis posterior.

- La tabla coaches fue limpiada con el propósito de responder preguntas relacionadas con los entrenadores, como identificar a los más exitosos y sus asociaciones con equipos específicos. Se seleccionaron las columnas coachID, firstName, lastName, tmID y won debido a su relevancia en las Preguntas 5 y 7.
  La columna coachID se utilizó como identificador único para los entrenadores, esencial para cruzar esta tabla con otras y asegurar la integridad de los datos. Las columnas firstName y lastName permiten mostrar los nombres completos de los entrenadores en los resultados. tmID se incluyó para identificar a los equipos con los que los entrenadores estuvieron asociados, mientras que la columna won fue fundamental para calcular las victorias y filtrar a los entrenadores más exitosos.
  Durante la limpieza, se normalizó el formato de coachID y tmID eliminando espacios y convirtiendo los valores a minúsculas, lo que asegura la uniformidad en los datos. Se eliminaron registros con valores nulos en coachID y tmID, ya que son claves importantes para el análisis. También se filtraron registros con valores negativos en won o lost, evitando inconsistencias. Además, se eliminaron duplicados exactos por la combinación de coachID, tmID y year para garantizar que no existieran conflictos en los datos.

- La tabla awards_coaches fue limpiada con el objetivo de responder preguntas relacionadas con los premios obtenidos por los entrenadores, específicamente la Pregunta 5.
  Se seleccionaron las columnas coachID, year y award debido a su relevancia. La columna coachID es clave para identificar a los entrenadores y cruzar esta tabla con la información detallada de los mismos, mientras que year y award son fundamentales para analizar qué premios se otorgaron y en qué años.
  Durante la limpieza, se normalizó el formato de coachID, eliminando espacios y convirtiendo los valores a minúsculas para garantizar la uniformidad y consistencia de los datos. Se eliminaron registros con valores nulos en coachID y award, ya que son necesarios para el análisis. La columna award fue normalizada para estandarizar los nombres de los premios. Además, se eliminaron duplicados exactos por la combinación de coachID, year y award para evitar contar múltiples veces el mismo premio para un entrenador en un año específico.
  Finalmente, se verificó la unicidad de coachID y se revisaron los premios únicos presentes en la tabla.

- Las tablas abbrev, awards_players y series_post no fueron limpiadas porque no son necesarias para responder las preguntas planteadas. Estas tablas contienen información que no se relaciona directamente con las métricas o relaciones analizadas en este trabajo. Si fuera necesario en un análisis futuro, se podrían procesar.

3.8 Modelo elegido

El modelo estrella fue seleccionado para estructurar los datos debido a su simplicidad, eficiencia en consultas y adaptabilidad. Este modelo organiza la información en una tabla central de hechos rodeada por varias tablas de dimensiones, lo que permite realizar análisis orientados al negocio de manera clara y efectiva. A continuación, se explican los principales motivos detrás de esta elección:

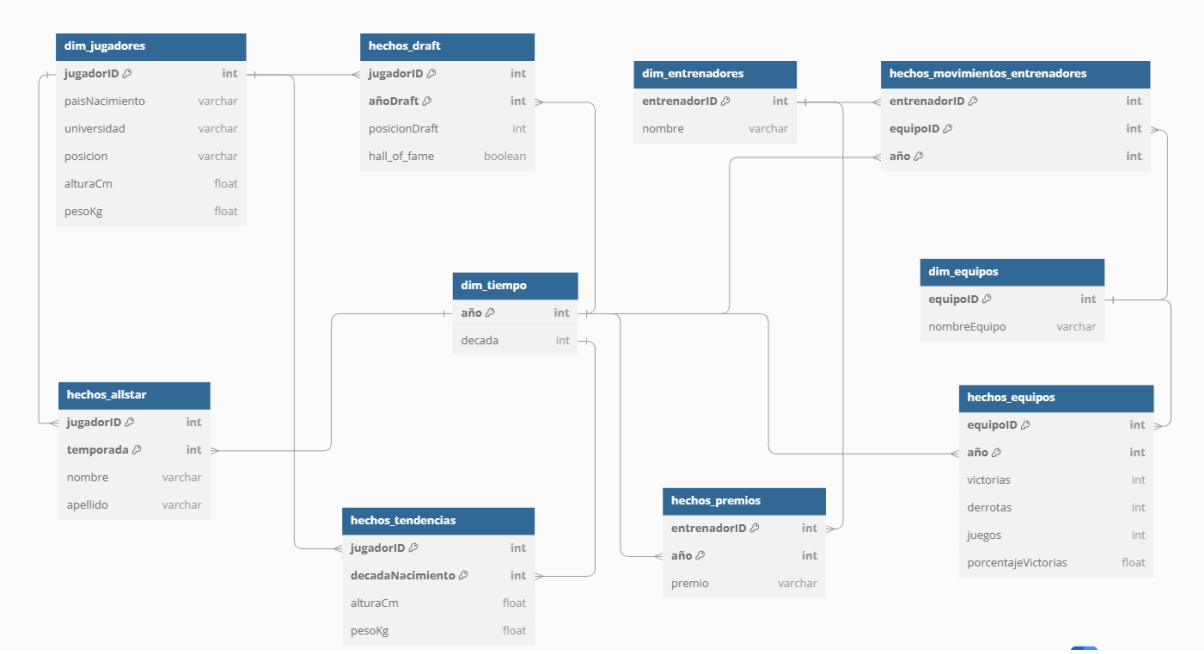
- **Simplicidad y consultas rápidas**:
  Las relaciones claras entre hechos y dimensiones minimizan la complejidad en las consultas, lo que permite responder preguntas como: "¿Cómo ha evolucionado la altura promedio de los jugadores por década?" o "¿Cuáles son los entrenadores con más premios?"
- **Eficiencia en el análisis de datos históricos**:
  Este modelo permite sumar, agrupar y analizar datos históricos clave, como victorias por temporada o posiciones en el Draft.
- **Flexibilidad para ampliar el modelo**:
  Nuevas métricas, preguntas analíticas o dimensiones pueden añadirse fácilmente sin afectar la estructura existente.
- **Adaptación a los datos procesados**:
  Nuevas métricas, preguntas analíticas o dimensiones pueden añadirse fácilmente sin afectar la estructura existente.

A continuación, se describe cada tabla de dimensiones y hechos en el modelo, junto con su propósito:

- **Tablas de Dimensiones**
  - **Dimensión de Jugadores**
    Esta tabla incluye atributos como el país de nacimiento, la universidad, la posición, la altura y el peso. Su clave primaria, jugadorID, facilita la conexión con las tablas de hechos, permitiendo análisis demográficos, físicos y de desempeño.
  - **Dimensión de Equipos**
    Contiene el identificador equipoID y el nombre de los equipos, lo que permite analizar métricas de rendimiento como victorias, derrotas y juegos.
  - **Dimensión de Entrenadores**
    Almacena el nombre y el identificador único entrenadorID, proporcionando información sobre los entrenadores para analizar su impacto en los equipos.
  - **Dimensión de Tiempo**
    Almacena los años y décadas, permitiendo realizar análisis temporales como

la evolución de las posiciones del Draft o el desempeño de equipos por temporada.

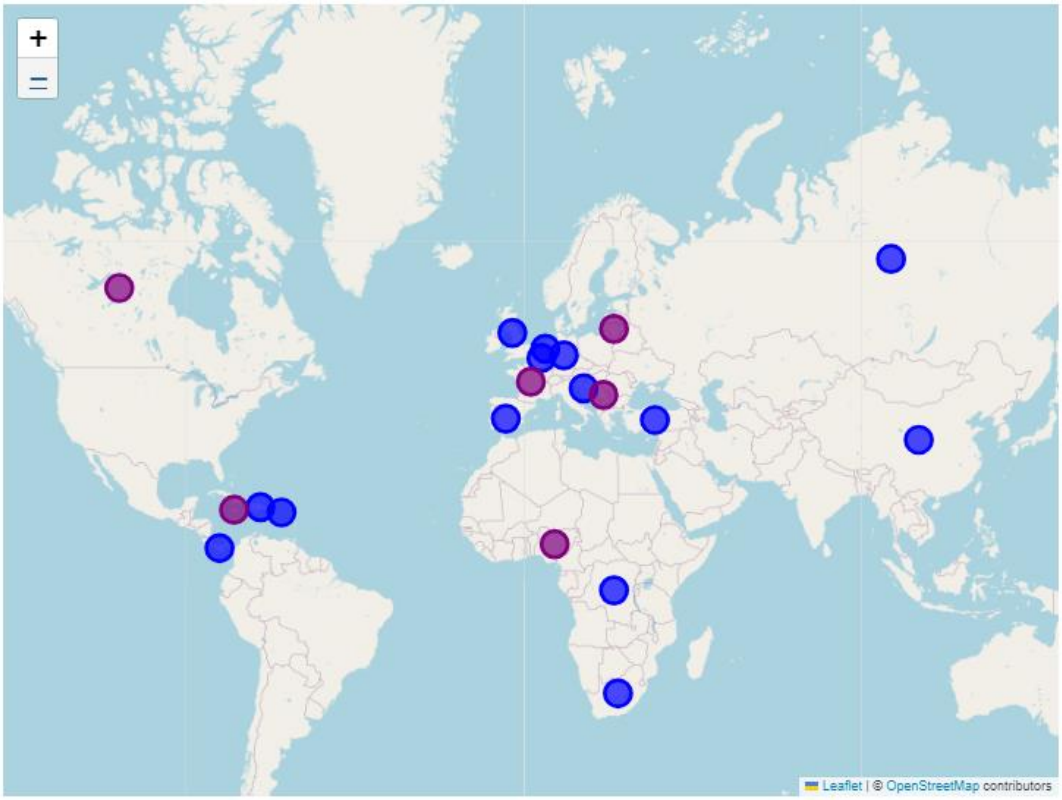
- **Tablas de Hechos**
  - **Hechos de Equipos**
    Registra métricas como victorias, derrotas y porcentaje de victorias por temporada. Responde preguntas como: "¿Cuáles son los equipos con las mejores temporadas en términos de victorias?"
  - **Hechos de Movimientos de Entrenadores**
    Rastrean cambios de equipos de los entrenadores a lo largo del tiempo, facilitando análisis sobre su trayectoria profesional.
  - **Hechos de Draft**
    Contienen datos sobre la posición en el Draft, el año de selección y si el jugador pertenece al Hall of Fame. Permiten analizar la importancia del Draft en la formación de equipos exitosos.
  - **Hechos de All-Star**
    Registran la participación de los jugadores en eventos All-Star, ayudando a identificar el impacto de los jugadores más destacados.
  - **Hechos de Premios**
    Proveen información sobre premios otorgados a entrenadores, ayudando a evaluar a los más exitosos.
  - **Hechos de Tendencias de Altura y Peso**
    Organizan datos físicos de jugadores por década de nacimiento, permitiendo estudiar tendencias y cambios a lo largo del tiempo.



3.9 Visualización de los resultados del análisis

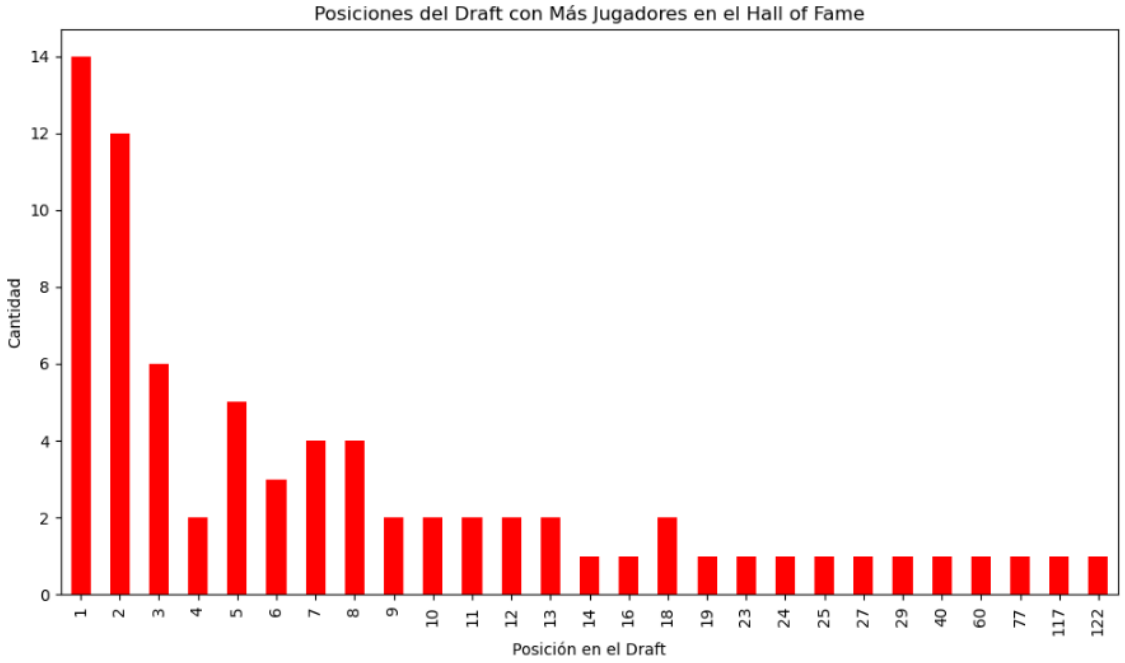
- **Pregunta 1: ¿Cuáles son los jugadores nacidos fuera de Estados Unidos que llegaron al All-Star (jugando al menos un minuto) o al Hall of Fame?**

[22]:



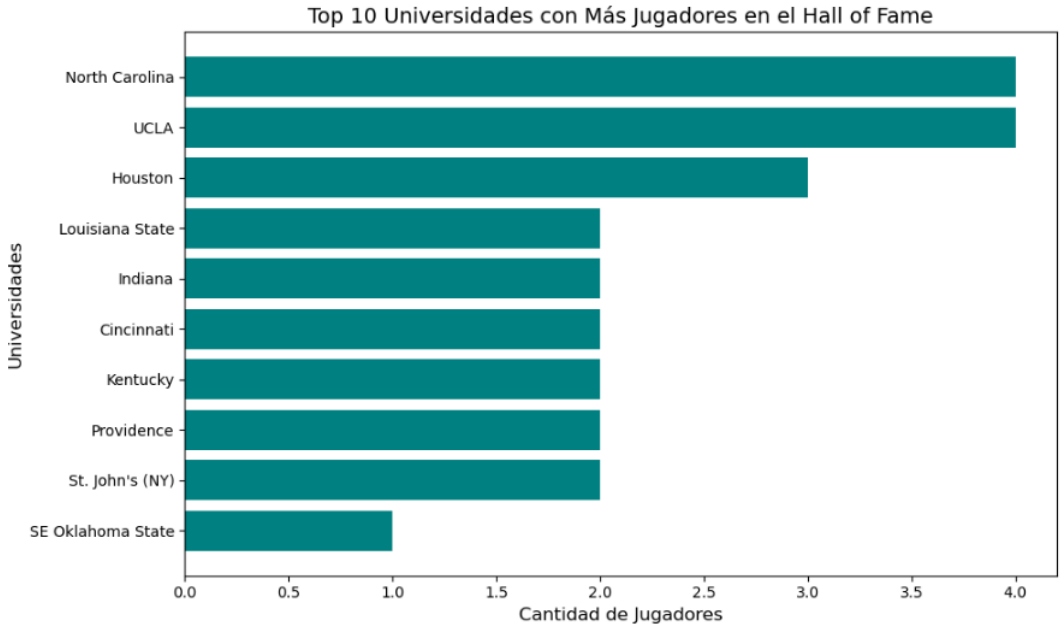
El análisis muestra que Europa domina como la principal región exportadora de talento hacia la NBA, con países como Alemania, Francia y Lituania destacándose. Canadá y Nigeria también sobresalen al tener jugadores en ambos grupos (All-Star y Hall of Fame). Regiones como América Latina, Asia y Oceanía presentan una representación limitada, reflejando posibles desigualdades en el desarrollo del baloncesto a nivel global.

- **Pregunta 2: ¿Cuáles son las posiciones del Draft que tienen más jugadores en el Hall of Fame?**



La mayoría de los jugadores en el Hall of Fame fueron seleccionados en las primeras posiciones del Draft, especialmente en el puesto 1, seguido por el 2 y el 3. Esto evidencia que las elecciones iniciales del Draft tienden a identificar a los talentos más destacados, aunque también hay casos excepcionales de jugadores seleccionados en posiciones más bajas que lograron destacar.

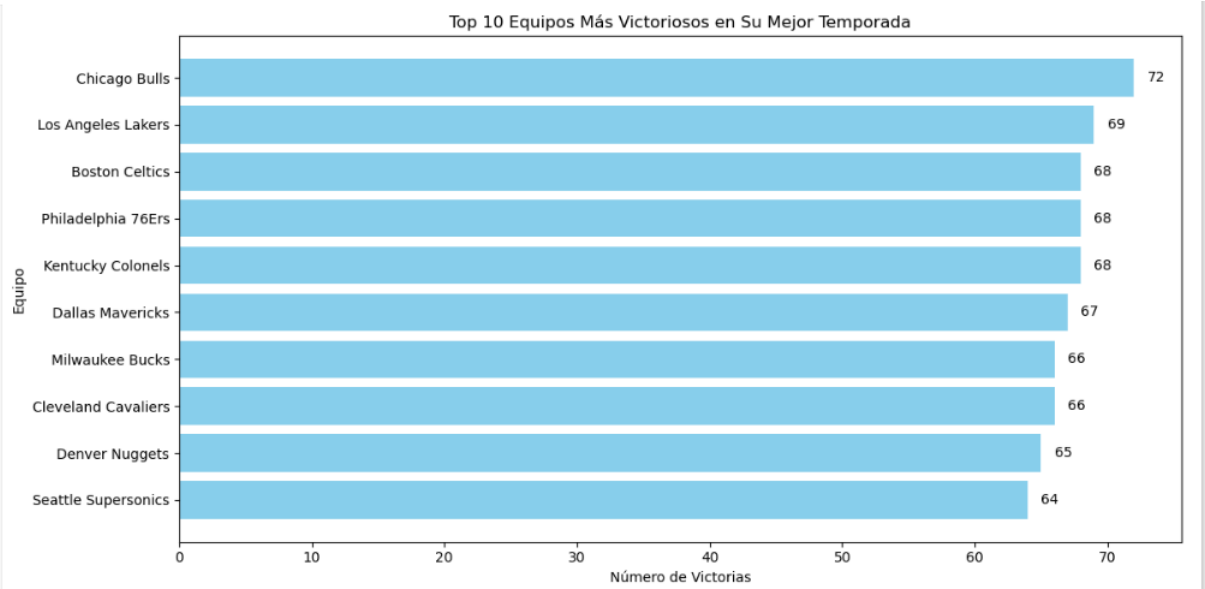
- **Pregunta 3:¿Cuáles son las 10 universidades que han producido la mayor cantidad de jugadores en el Hall of Fame?**



La gráfica muestra que las universidades North Carolina y UCLA son las principales instituciones académicas en la formación de jugadores que han llegado al Hall of Fame, con una representación notablemente mayor en comparación con otras universidades. Esto

refleja su éxito histórico en el desarrollo de talentos destacados para la NBA. Universidades como Houston, Louisiana State, e Indiana también tienen un impacto considerable, aunque menor. La presencia de universidades menos conocidas, como SE Oklahoma State, sugiere que el talento excepcional puede surgir de diversas instituciones, incluso aquellas con menor reconocimiento en el ámbito deportivo. Este análisis pone de manifiesto el predominio de ciertas universidades en la formación de élite, aunque deja espacio para excepciones notables.

- **Pregunta 4: "¿Cuáles son los 10 equipos con las mejores temporadas en términos de número de victorias?"**

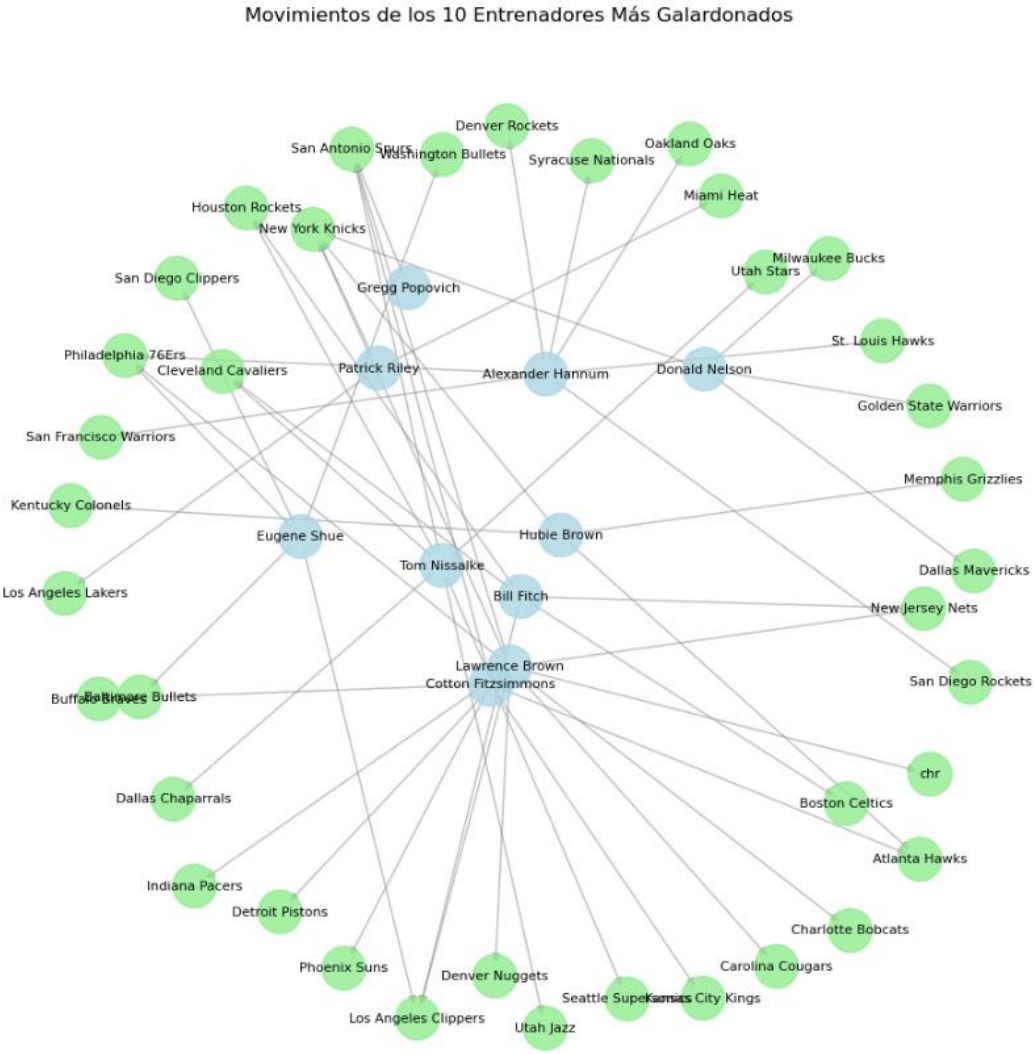


La gráfica muestra que los Chicago Bulls lideran como el equipo con más victorias en su mejor temporada, alcanzando 72 triunfos, seguidos por los Los Angeles Lakers con 69 victorias. Equipos históricos como los Boston Celtics, Philadelphia 76ers y Kentucky Colonels también destacan con 68 victorias cada uno. Esto refleja la dominancia de estos equipos en sus mejores campañas, consolidando su legado en la NBA.
El desempeño destacado de otros equipos, como los Dallas Mavericks, Milwaukee Bucks, y Cleveland Cavaliers, demuestra que el éxito en la NBA no se limita a las franquicias más reconocidas históricamente, sino que también hay temporadas excepcionales de equipos emergentes. Este análisis subraya la competitividad de la liga y la capacidad de diferentes equipos para alcanzar la excelencia en momentos clave de su historia.

- **Pregunta 5: ¿Quiénes son los entrenadores con más premios, y cómo se distribuyen los premios entre ellos?**

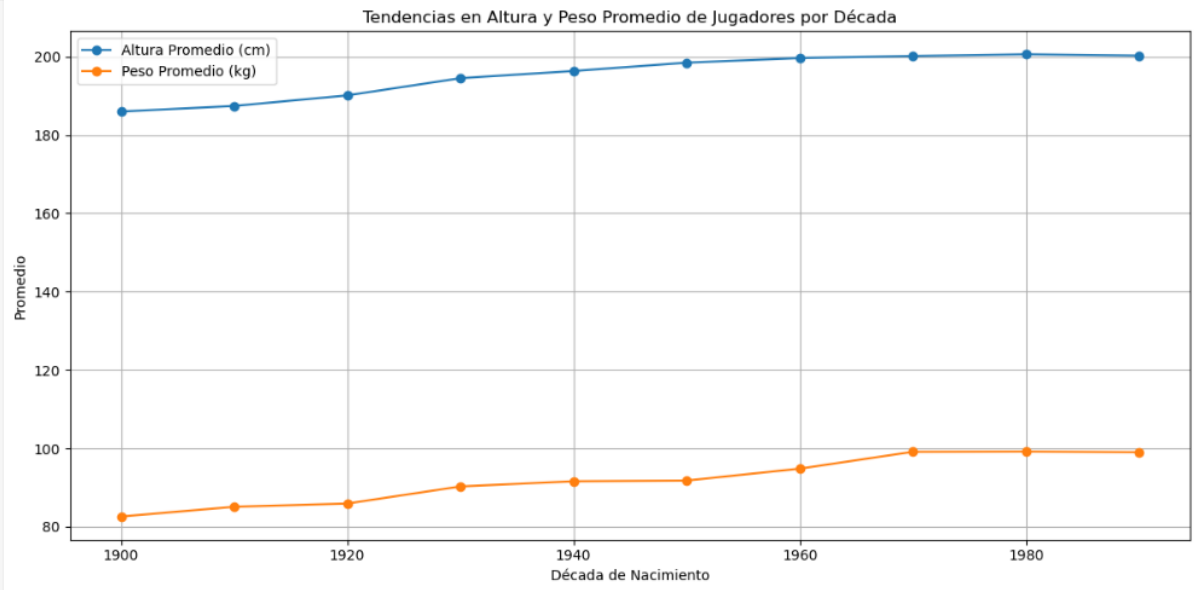
La gráfica muestra que Lawrence Brown lidera como el entrenador más galardonado, acumulando el 16.7% de los premios entre los entrenadores destacados. Le siguen Donald Nelson y Patrick Riley con el 12.5% cada uno. El resto de los premios están distribuidos equitativamente entre otros entrenadores como Cotton Fitzsimmons, Gregg Popovich, y Hubie Brown, cada uno con un 8.33%. Esto refleja que, aunque hay una ligera concentración de premios en unos pocos entrenadores, existe una diversidad en el reconocimiento de los logros en la liga.

- **Pregunta 6 ¿Cuáles fueron los movimientos de los 10 entrenadores más galardonados entre equipos a lo largo de su carrera?**



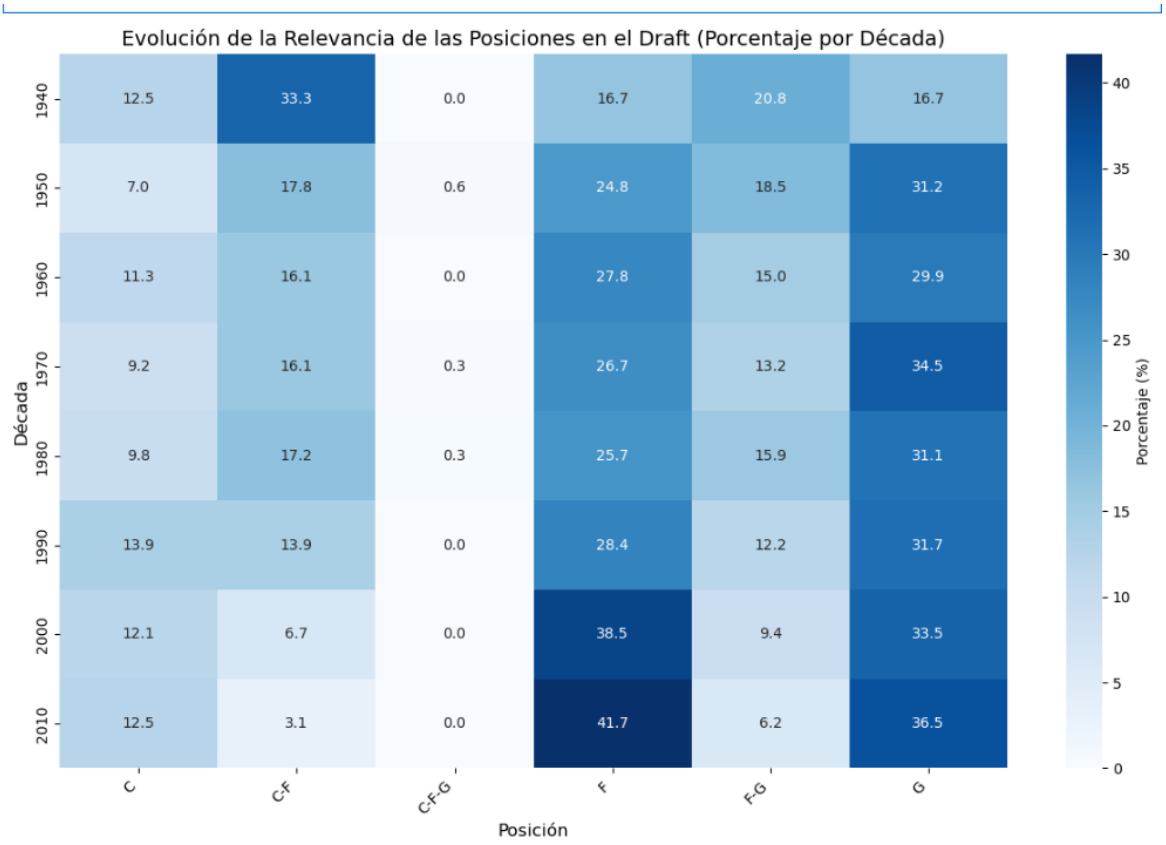
La red de movimientos muestra que los entrenadores más galardonados han dirigido múltiples equipos a lo largo de sus carreras, con algunos de ellos, como Lawrence Brown y Hubie Brown, estableciendo conexiones con una gran cantidad de franquicias. Esto evidencia la movilidad y la influencia de estos entrenadores dentro de la NBA, reflejando cómo su experiencia ha sido valorada en distintos contextos y equipos. Además, destaca la interacción entre equipos históricos, mostrando patrones de colaboración y cambio dentro de la liga.

- **Pregunta 7 ¿Cómo han evolucionado la altura y el peso promedio de los jugadores a lo largo de las décadas?**



El análisis muestra que tanto la altura como el peso promedio de los jugadores han aumentado de forma constante a lo largo de las décadas, especialmente en las primeras del siglo XX. Sin embargo, a partir de las décadas de 1970 y 1980, estas tendencias comienzan a estabilizarse, sugiriendo un límite en las características físicas promedio de los jugadores. Esto refleja la evolución del baloncesto hacia un deporte con mayores demandas físicas, aunque también podría indicar que se ha alcanzado un punto de equilibrio en el desarrollo físico de los atletas.

- **Pregunta 8: ¿Cómo ha cambiado la relevancia de cada posición en el Draft a lo largo de las décadas?**



El análisis muestra cómo ha cambiado la relevancia de las diferentes posiciones en el Draft de la NBA a lo largo de las décadas. Se observa que las posiciones de Forward (F) y Guard (G) han sido consistentemente dominantes, manteniendo una alta proporción de selecciones en la mayoría de las décadas. Sin embargo, las posiciones híbridas, como Center-Forward (C-F) o Forward-Guard (F-G), han tenido poca relevancia relativa, lo que sugiere que los equipos suelen priorizar jugadores con roles más definidos. Además, la posición de Center (C) ha visto una leve disminución en su importancia en las décadas más recientes, posiblemente reflejando un cambio en las estrategias de juego hacia un enfoque más versátil y orientado al perímetro.

3.10 Conclusiones

A partir del análisis realizado, se pueden extraer conclusiones generales sobre los patrones y tendencias en la NBA a lo largo del tiempo. Se evidencia que ciertos aspectos, como las posiciones en el Draft, las contribuciones de universidades específicas y las trayectorias de entrenadores y equipos, han evolucionado significativamente, reflejando cambios en las estrategias, la globalización del deporte y el desarrollo del baloncesto como disciplina.

La preeminencia de Europa como principal exportador de talento hacia la NBA refuerza la idea de que la liga es un escenario global, aunque persisten desigualdades en la representación de otras regiones como América Latina y Asia. Por otro lado, la tendencia hacia jugadores más altos y pesados a lo largo de las décadas muestra cómo la evolución física ha respondido a las demandas del juego moderno.

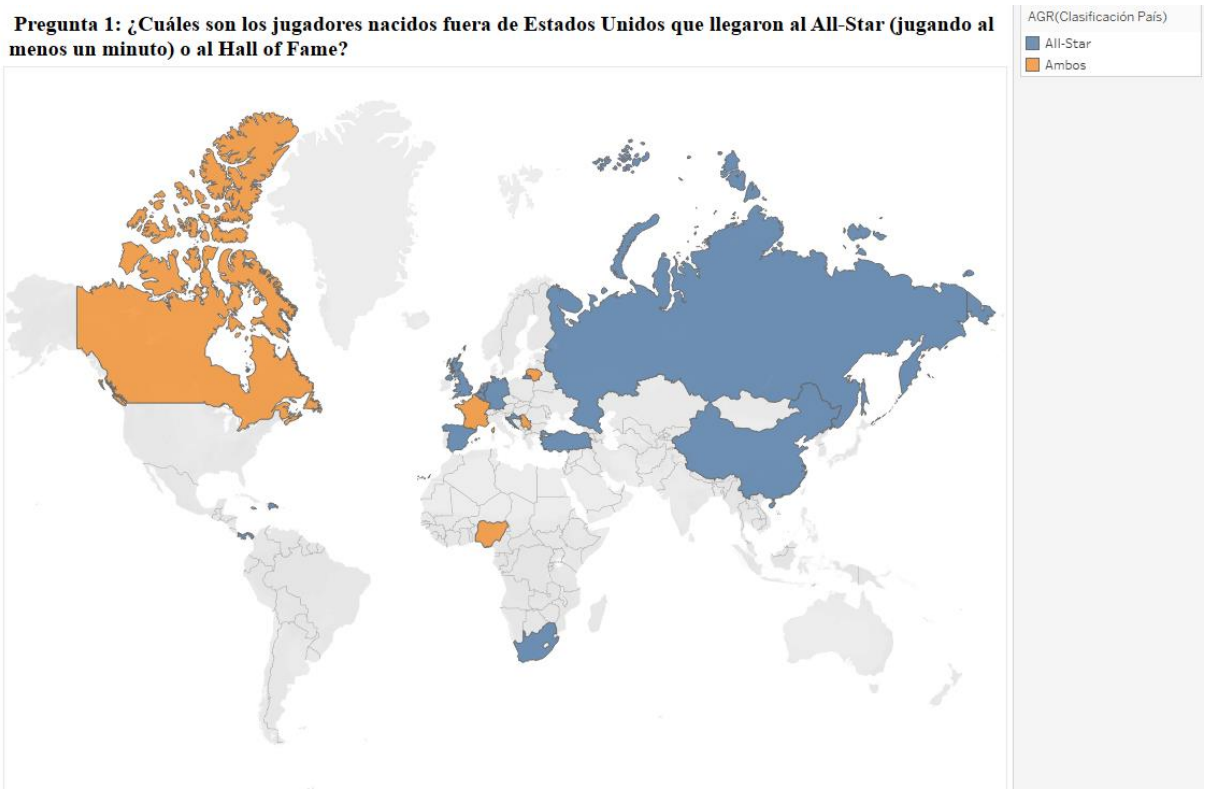
En conjunto, estos hallazgos resaltan no solo la rica historia de la liga, sino también la dinámica cambiante que continúa definiendo el baloncesto profesional.

4. Tableau

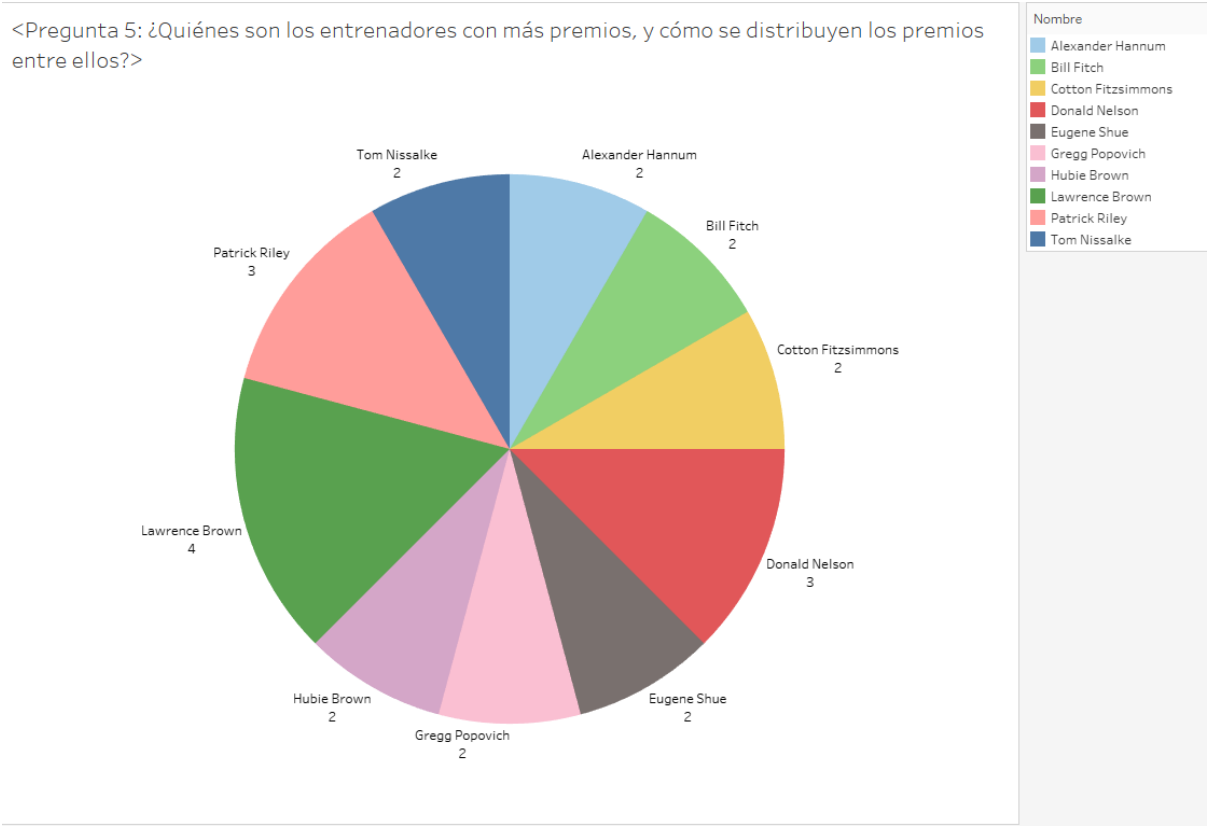
Para realizar las visualizaciones en Tableau, se generaron y guardaron las siguientes tablas en la carpeta "Anl", cada una diseñada para responder preguntas específicas:

1. **posiciones_hall_of_fame.csv**: Contiene la cantidad de jugadores en el Hall of Fame agrupados por su posición en el Draft.
2. **entrenadores_premios.csv**: Incluye el total de premios ganados por cada entrenador, ordenados por cantidad de premios.
3. **top_equipos_victorias.csv**: Muestra los equipos con el mayor número de victorias en cualquier temporada, destacando los 10 mejores.
4. **jugadores_allstar_hof_mejorado.csv**: Filtra jugadores nacidos fuera de Estados Unidos que participaron en el All-Star o están en el Hall of Fame, categorizándolos según su participación en ambos.

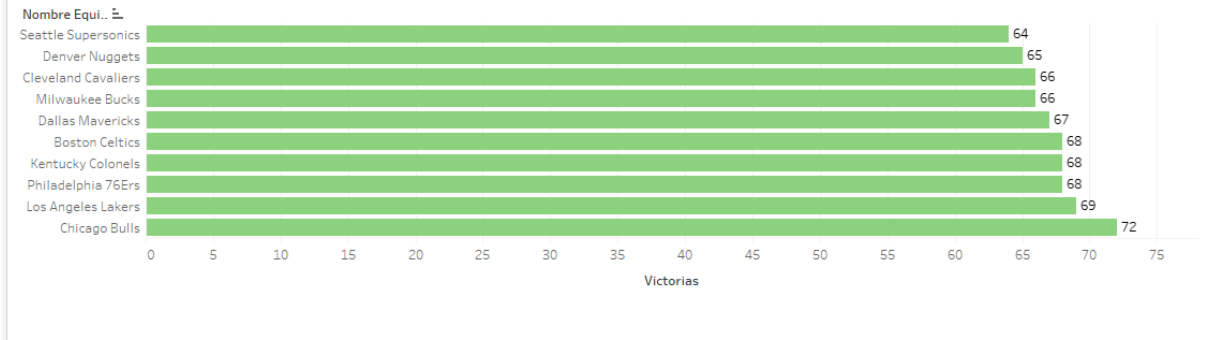
Estas tablas fueron exportadas como archivos CSV, facilitando su integración con Tableau para generar visualizaciones claras y comprensibles.



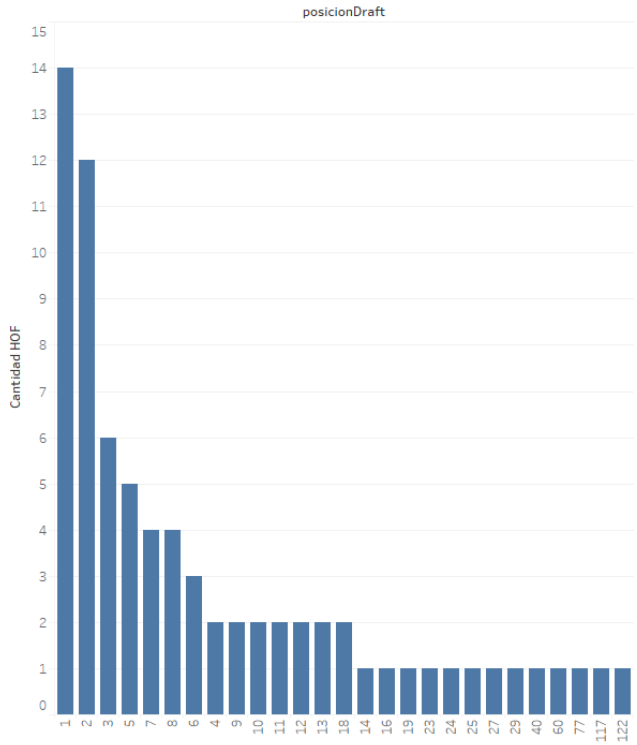
<Pregunta 5: ¿Quiénes son los entrenadores con más premios, y cómo se distribuyen los premios entre ellos?>



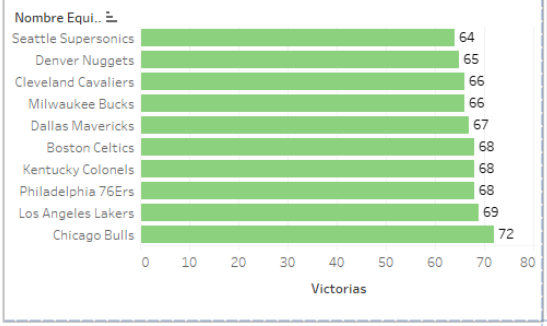
<Pregunta 4: ¿Cuáles son los 10 equipos con las mejores temporadas en términos de número de victorias?>



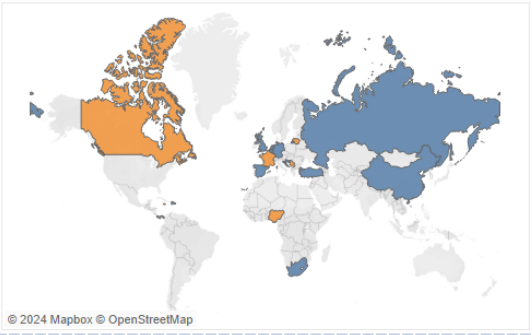
<Pregunta 2: ¿Cuáles son las posiciones del Draft que tienen más jugadores en el Hall of Fame?>



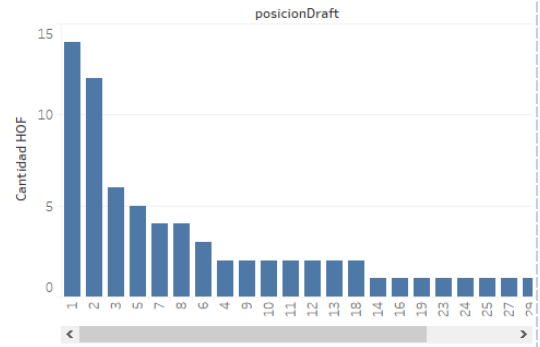
<Pregunta 4: ¿Cuáles son los 10 equipos con las mejores temporadas en términos de número de victorias?>



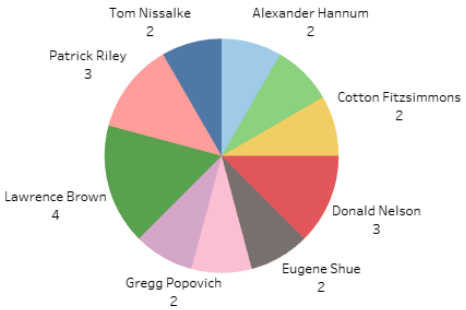
Pregunta 1: ¿Cuáles son los jugadores nacidos fuera de Estados Unidos que llegaron al All-Star (jugando al menos un minuto) o al Hall of Fame?



<Pregunta 2: ¿Cuáles son las posiciones del Draft que tienen más jugadores en el Hall of Fame?>



<Pregunta 5: ¿Quiénes son los entrenadores con más premios, y cómo se distribuyen los premios entre ellos?>



5. Comparación de la arquitectura de un datalake con otras soluciones

5.1. Introducción

Un datalake es un repositorio centralizado que permite almacenar todos los datos estructurados y no estructurados a cualquier escala. En su implementación tradicional, se suelen utilizar herramientas como Apache Hive, Apache Spark y Apache NiFi para gestionar, procesar y orquestar los datos. Sin embargo, hoy en día, las principales plataformas en la nube como Amazon Web Services (AWS), Azure y Google Cloud ofrecen soluciones que permiten construir un datalake similar utilizando sus servicios propios. En este análisis, examino cómo se pueden utilizar estas herramientas en dichas plataformas en la nube, destacando las opciones equivalentes para Hive, Spark, NiFi y otros componentes clave en la arquitectura de un datalake.

5.2. Almacenamiento

- **En Apache (Hadoop):** El almacenamiento en un datalake tradicionalmente utiliza HDFS (Hadoop Distributed File System) para almacenar grandes volúmenes de datos. HDFS permite el almacenamiento de datos no estructurados, semiestructurados y estructurados.
- **AWS (Amazon S3):** Amazon S3 es el servicio de almacenamiento en la nube de AWS que puede reemplazar HDFS. Ofrece almacenamiento escalable para datos en bruto y soporta acceso desde otros servicios como AWS Glue, Amazon EMR, etc. S3 es ideal para crear un datalake debido a su capacidad de manejar datos en su formato original sin necesidad de procesarlos previamente.
- **Azure (Azure Blob Storage):** En Azure, Blob Storage o Azure Data Lake Storage Gen2 son las soluciones utilizadas para almacenar datos no estructurados, semiestructurados y estructurados. Azure Data Lake Gen2 está especialmente optimizado para grandes volúmenes de datos y soporta HDFS de manera nativa, similar a HDFS en Hadoop.
- **Google Cloud (Google Cloud Storage):** Google Cloud Storage proporciona almacenamiento escalable similar a Amazon S3. Permite almacenar grandes volúmenes de datos sin procesarlos y es utilizado para implementar un datalake en Google Cloud.

Conclusión: Las tres plataformas ofrecen almacenamiento de datos en la nube que puede ser utilizado de manera similar al HDFS en Apache. Amazon S3, Azure Blob Storage y Google Cloud Storage son alternativas equivalentes para almacenar datos sin procesar en un datalake.

5.3. Procesamiento de Datos

- **En Apache (Hive y Spark):** Apache Hive se utiliza para consultas SQL sobre grandes volúmenes de datos en un datalake basado en HDFS, mientras que Apache Spark es utilizado para procesamiento en memoria de grandes datos en paralelo, especialmente útil para análisis en tiempo real.
- **AWS (Amazon EMR):** Amazon EMR es un servicio que permite ejecutar Apache Spark, Apache Hive, Hadoop y otras herramientas de procesamiento de datos de manera escalable. Ofrece un entorno gestionado para ejecutar grandes cargas de

trabajo de procesamiento de datos sin necesidad de configurar clústeres manualmente.

- **Azure (Azure Databricks y HDInsight):** Azure Databricks es una plataforma optimizada para ejecutar Apache Spark y está integrada con otros servicios de Azure para análisis de grandes volúmenes de datos. HDInsight es otro servicio de Azure que permite ejecutar Apache Hadoop, Hive y Spark de forma gestionada en la nube.
- **Google Cloud (Google Cloud Dataproc):** Google Cloud Dataproc es un servicio gestionado que permite ejecutar Apache Hadoop, Apache Spark y Apache Hive en la nube. Al igual que Amazon EMR y Azure Databricks, es ideal para ejecutar procesos de análisis de grandes volúmenes de datos.

Conclusión: En cuanto al procesamiento, las tres plataformas ofrecen soluciones gestionadas que permiten ejecutar herramientas de Apache como Hive y Spark. Amazon EMR, Azure Databricks y Google Cloud Dataproc facilitan la ejecución de trabajos de procesamiento de datos sin necesidad de manejar clústeres manualmente.

5.4. Orquestación de Datos

- **En Apache (NiFi):** Apache NiFi se utiliza para orquestar flujos de datos entre sistemas y gestionar la integración de datos en un datalake. Facilita la transferencia, transformación y enrutamiento de datos entre diferentes sistemas.
- **AWS (AWS Glue):** AWS Glue es un servicio totalmente gestionado de ETL (Extracción, Transformación y Carga) que facilita la integración de datos en un datalake de AWS. Glue gestiona la orquestación de flujos de datos y permite ejecutar trabajos de transformación y carga de datos de forma eficiente.
- **Azure (Azure Data Factory):** Azure Data Factory es un servicio de integración de datos y orquestación de flujos de trabajo en Azure. Permite la orquestación de procesos de transformación de datos, similares a lo que hace Apache NiFi, pero en la nube de Azure.
- **Google Cloud (Google Cloud Dataflow):** Google Cloud Dataflow es un servicio de procesamiento de datos en tiempo real basado en Apache Beam. Facilita la integración y orquestación de datos en Google Cloud, similar a cómo lo hace Apache NiFi en un entorno local.

Conclusión: Para la orquestación de datos, las tres plataformas ofrecen servicios que son equivalentes a Apache NiFi. AWS Glue, Azure Data Factory y Google Cloud Dataflow permiten gestionar flujos de datos y realizar tareas ETL de forma eficiente en la nube.

5.5. Seguridad y Gobernanza de Datos

- **En Apache (HDFS y Hive):** En un datalake basado en Apache, la seguridad se gestiona mediante herramientas como Apache Ranger o Apache Sentry, que permiten controlar el acceso a los datos almacenados en HDFS y las consultas realizadas en Hive.
- **AWS (AWS Lake Formation e IAM):** AWS Lake Formation permite configurar la seguridad y la gobernanza de datos en un datalake de Amazon, mientras que AWS IAM (Identity and Access Management) permite gestionar los permisos y accesos a los recursos en AWS. Estos servicios ayudan a implementar políticas de seguridad en el datalake.

- **Azure (Azure Active Directory y Azure Purview):** Azure Active Directory (AAD) es utilizado para gestionar el acceso a los recursos en Azure, mientras que Azure Purview proporciona gobernanza de datos, clasificación y control de acceso sobre los datos almacenados en Azure.
- **Google Cloud (Google Cloud IAM y Data Catalog):** Google Cloud IAM permite gestionar los permisos de acceso a los recursos en Google Cloud, mientras que Google Cloud Data Catalog ofrece herramientas para la gobernanza de datos, permitiendo clasificar, etiquetar y gestionar el acceso a los datos en Google Cloud.

Conclusión: Las tres plataformas ofrecen herramientas equivalentes a Apache Ranger y Apache Sentry para la gestión de seguridad y gobernanza de datos. AWS Lake Formation, Azure Purview y Google Cloud Data Catalog son las soluciones que permiten gestionar la seguridad y la gobernanza de los datos en la nube.

5.6. Conclusión General

En resumen, he analizado cómo AWS, Azure y Google Cloud ofrecen soluciones equivalentes para almacenar, procesar, orquestar y gobernar los datos en un datalake. Cada plataforma tiene sus herramientas propias, pero los servicios principales son similares en todas ellas:

- **Almacenamiento:** S3 (AWS), Blob Storage (Azure), Cloud Storage (Google)
- **Procesamiento:** EMR (AWS), Databricks (Azure), Dataproc (Google)
- **Orquestación:** Glue (AWS), Data Factory (Azure), Dataflow (Google)
- **Seguridad y Gobernanza:** IAM y Lake Formation (AWS), Active Directory y Purview (Azure), IAM y Data Catalog (Google)

Cada plataforma tiene sus características específicas, pero todas ofrecen una infraestructura escalable para construir un datalake en la nube.

6. Reflexión y aprendizajes

Trabajar con los mismos datos tanto en Pandas como en Spark fue una experiencia enriquecedora. Esto me permitió comparar las fortalezas de cada herramienta: mientras que Pandas es ideal para manejar conjuntos de datos pequeños o medianos y facilita el análisis exploratorio localmente, Spark destacó por su capacidad para procesar grandes volúmenes de datos de manera distribuida y eficiente.

Un aspecto fundamental del trabajo fue el modelado de los datos, que implicó organizar la información en un esquema que permitiera responder preguntas de manera clara y estructurada. Este paso me ayudó a comprender la importancia de un diseño adecuado para facilitar análisis eficientes y garantizar la consistencia en los resultados.

También enfrenté el desafío de formular preguntas para guiar el análisis. Aunque las preguntas seleccionadas fueron relevantes para los datos, reconozco que algunas no aprovecharon completamente el enfoque de Big Data. Esto me dejó como aprendizaje la necesidad de enfocar mejor las preguntas hacia patrones más complejos o análisis que resalten el potencial de las herramientas utilizadas.

Por último, aprendí a utilizar Tableau y herramientas de visualización en Pandas, lo que me permitió presentar los resultados de forma clara y efectiva. Estas herramientas facilitaron comunicar los hallazgos de manera visual y comprensible, conectando los análisis con respuestas concretas para las preguntas planteadas.

7. Apéndice: Uso de inteligencia artificial.

Durante el desarrollo de este obligatorio, utilicé herramientas de inteligencia artificial para optimizar algunas etapas del trabajo. En particular, recurrí a un asistente de IA para:

1. **Redacción de textos:** Solicité sugerencias para estructurar secciones del informe, como la descripción de tablas, reflexiones y aprendizajes. La IA me ayudó a organizar ideas y redactar textos más claros y completos.
2. **Formulación de comparaciones:** Consulté sobre cómo comparar los resultados obtenidos entre Pandas y Spark, asegurándome de que los textos reflejaran de manera precisa las similitudes y diferencias entre ambas herramientas.
3. **Resolución de dudas técnicas:** La IA me apoyó con explicaciones sobre conceptos relacionados con Big Data, como el modelado de datos, la preparación de información para análisis y visualizaciones, y cómo realizar consultas específicas en Pandas para generar visualizaciones relevantes.

Es importante destacar que, aunque utilicé la IA como apoyo, todas las decisiones finales sobre qué incluir en el informe, las consultas realizadas y la interpretación de los resultados fueron tomadas por mí. Esto asegura que el trabajo refleje mis aprendizajes y análisis personales.