

Universidad ORT Uruguay

Facultad de Ingeniería

Escuela de Tecnología

DOCUMENTO DE ANÁLISIS

Obligatorio de Big Data

Emilio Colina– 229705

Guillermo Montecoral – 259620

Profesor: Eduardo García

Grupo: N6A

Analista en Tecnología de la Información

01/12/2025

Contenido del documento

1. Caso de estudio.....	3
2. Problema planteado.....	4
3. Análisis exploratorio.....	5
3.1. Análisis de datos.....	5
3.2. Carga de datos	5
3.3. Dimensiones y columnas	5
3.4. Descripción de cada archivo.....	6
3.5. Control de valores nulos	9
3.6. Valores duplicados por columnas.....	11
3.7. Valores duplicados por filas.....	12
3.8. Modelo de datos normalizado.....	12
3.9. Limpieza de los archivos	12
3.10. Resumen de limpieza y preparación de datos.....	18
4. Estructura del modelo para Hive.....	20
5. Desarrollo de análisis.....	21
7. Dashboard en Power BI	32
8. Conclusiones	33
9. Uso de inteligencia artificial	34

1. Caso de estudio

El propósito de este trabajo es procesar, filtrar y analizar datos históricos del fútbol inglés utilizando herramientas y enfoques propios del Big Data. Para ello, se emplearon tecnologías como **Python**, aprovechando **Pandas** en las etapas iniciales de exploración y limpieza de los datos, así como también **Apache Spark**, con el fin de estudiar la evolución estructural y competitiva de sus ligas a lo largo del tiempo. El objetivo principal es ofrecer una visión integral que integre aspectos deportivos, espaciales y evolutivos del sistema futbolístico inglés.

El dataset empleado fue proporcionado por la Facultad ORT y posteriormente enriquecido con información georreferenciada incorporada por nuestro equipo. Este conjunto de datos abarca información detallada de todas las divisiones del fútbol inglés desde 1888 hasta 2021, incluyendo datos a nivel de partido y variables específicas de cada club, como el año de su primera participación en competencias oficiales.

2. Problema planteado

A partir del conjunto de datos surgen diversas interrogantes que permiten explorar la evolución competitiva y territorial del deporte a lo largo del tiempo.

Con base en esta información, se plantean las siguientes preguntas de análisis:

1. **Analizar la evolución de la cantidad de partidos por temporada**, desglosando los resultados en empates y victorias, con el fin de identificar patrones competitivos a lo largo del tiempo.
2. **Examinar la evolución histórica de las distintas categorías del fútbol inglés**, considerando su periodo de existencia y los cambios estructurales en el sistema de ligas.
3. **Estudiar la variación en la cantidad de equipos participantes en la máxima categoría**, evaluando cómo se ha modificado su composición desde sus inicios.
4. **Explorar la distribución de equipos según sus victorias como visitantes y derrotas como locales**, para identificar comportamientos competitivos relevantes.
5. **Determinar geográficamente la concentración de los clubes campeones de la máxima categoría**, con el fin de visualizar patrones espaciales en la historia del fútbol inglés.
6. **Analizar la frecuencia y el ritmo de crecimiento en la cantidad de equipos que han participado en las diferentes divisiones del fútbol inglés**, identificando etapas de expansión, reestructuración y consolidación del sistema de ligas.
7. **Calcular el porcentaje promedio de victorias de los clubes campeones de la máxima categoría desde 1980**, para evaluar el nivel de dominio competitivo requerido para obtener el título en distintos periodos históricos.
8. **En que temporada se comienza a conceder 3 puntos por partidos ganados**, analizando el cambio normativo en el contexto de la evolución competitiva del fútbol inglés.

3. Análisis exploratorio

3.1. Análisis de datos

El análisis se llevó a cabo utilizando dos herramientas con el objetivo de practicar y comprender el funcionamiento de cada una: **Pandas** y **Apache Spark**.

La herramienta principal utilizada fue **Pandas**, que resultó fundamental para manipular, comprender y explorar los datos proporcionados. Con Pandas pudimos realizar operaciones como filtrado de registros, inspección de tipos de datos por columna, conteo de columnas y filas, así como agrupar, transformar o eliminar columnas y filas según fuera necesario. Estas funcionalidades nos permitieron manejar y procesar grandes volúmenes de datos de manera eficiente, facilitando un análisis más detallado y estructurado.

De forma complementaria, también se empleó **Apache Spark**, especialmente para practicar su enfoque distribuido y validar los resultados obtenidos con Pandas. Fue útil para verificar la consistencia de los datos, realizar exploraciones iniciales y comprender cómo se comportan las mismas transformaciones en un motor distribuido.

3.2. Carga de datos

Los datos fueron cargados desde archivos CSV, aplicando un proceso de normalización y limpieza inicial.

Durante la carga se manejaron posibles errores y se estandarizaron los nombres de las columnas, convirtiéndolos a minúsculas, eliminando espacios en blanco y reemplazando cualquier espacio dentro de un nombre de columna por un guion bajo (_).

Se identificaron valores nulos y se controlaron posibles inconsistencias en los archivos, como columnas con más de un tipo de dato.

Para facilitar la manipulación y exploración de los datos, todos los dataframes fueron almacenados en un diccionario, lo que permitió un acceso más flexible y simplificó la integración de la información durante el análisis.

3.3. Dimensiones y columnas

Antes de aplicar procesos de limpieza, se realizó un análisis exploratorio de la estructura de los archivos originales.

En esta etapa se identificó la cantidad de registros por filas, la cantidad de atributos por columnas, los nombres de columnas y el tipo de datos presente en cada una.

Este procedimiento permitió comprender la estructura general y el tipo de información contenida en cada tabla. Las columnas representan los atributos que describen cada registro, y sus nombres brindan información sobre el tipo de datos que cada tabla almacena, facilitando así la planificación de transformaciones y análisis posteriores.

3.4. Descripción de cada archivo

Appearances

Descripción: Contiene información detallada de la participación de los equipos en cada partido, incluyendo resultados, goles y puntos obtenidos en cada temporada.

Dimensiones:

- Filas: **407.912**
- Columnas: **22**

Columnas:

- **key_id:** Identificador único de la fila
- **season_id:** Identificador de la temporada
- **season:** Año de la temporada
- **tier:** Nivel o categoría de la liga
- **division:** División del torneo
- **subdivision:** Subdivisión dentro de la división
- **match_id:** Identificador del partido
- **match_name:** Nombre del partido
- **team_id:** Identificador del equipo
- **team_name:** Nombre del equipo
- **opponent_id:** Identificador del equipo contrario
- **opponent_name:** Nombre del equipo contrario
- **home_team:** Indicador de equipo local
- **away_team:** Indicador de equipo visitante
- **goals_for:** Goles a favor del equipo
- **goals_against:** Goles en contra del equipo
- **goal_difference:** Diferencia de goles
- **result:** Resultado del partido
- **win:** Número de victorias
- **lose:** Número de derrotas
- **draw:** Número de empates
- **points:** Puntos obtenidos

Matches

Descripción: Contiene información sobre los partidos disputados, incluyendo equipos participantes, resultados y margen de goles.

Dimensiones:

- Filas: **203.956**
- Columnas: **21**

Columnas:

- **key_id**: Identificador único de la fila
- **season_id**: Identificador de la temporada
- **season**: Año de la temporada
- **tier**: Nivel o categoría de la liga
- **division**: División del torneo
- **subdivision**: Subdivisión dentro de la división
- **match_id**: Identificador del partido
- **match_name**: Nombre del partido
- **home_team_id**: Identificador del equipo local
- **home_team_name**: Nombre del equipo local
- **away_team_id**: Identificador del equipo visitante
- **away_team_name**: Nombre del equipo visitante
- **score**: Marcador final
- **home_team_score**: Goles del equipo local
- **away_team_score**: Goles del equipo visitante
- **home_team_score_margin**: Diferencia de goles del equipo local
- **away_team_score_margin**: Diferencia de goles del equipo visitante
- **result**: Resultado del partido
- **home_team_win**: Indica si el equipo local ganó
- **away_team_win**: Indica si el equipo visitante ganó
- **draw**: Indica empate

Seasons

Descripción: Contiene información general sobre cada temporada, incluyendo la división, el ganador y la cantidad de equipos participantes.

Dimensiones:

- Filas: **431**
- Columnas: **8**

Columnas:

- **key_id**: Identificador único de la fila
- **season_id**: Identificador de la temporada
- **season**: Año de la temporada
- **tier**: Nivel o categoría de la liga
- **division**: División del torneo

- **subdivision:** Subdivisión dentro de la división
- **winner:** Equipo ganador de la temporada
- **count_teams:** Cantidad de equipos participantes

Standings

Descripción: Contiene la clasificación de los equipos en cada temporada, incluyendo posiciones, partidos jugados, resultados y puntos.

Dimensiones:

- Filas: **9.556**
- Columnas: **18**

Columnas:

- **key_id:** Identificador único de la fila
- **season_id:** Identificador de la temporada
- **season:** Año de la temporada
- **tier:** Nivel o categoría de la liga
- **division:** División del torneo
- **subdivision:** Subdivisión dentro de la división
- **position:** Posición del equipo en la tabla
- **team_id:** Identificador del equipo
- **team_name:** Nombre del equipo
- **played:** Partidos jugados
- **wins:** Partidos ganados
- **draws:** Partidos empatados
- **losses:** Partidos perdidos
- **goals_for:** Goles a favor
- **goals_against:** Goles en contra
- **goal_difference:** Diferencia de goles
- **points:** Puntos obtenidos
- **point_adjustment:** Ajustes de puntos si aplica

Teams

Descripción: Contiene información sobre los equipos, incluyendo identificadores, nombres actuales y anteriores, y estado de actividad.

Dimensiones:

- Filas: **144**
- Columnas: **8**

Columnas:

- **key_id:** Identificador único de la fila
- **team_id:** Identificador del equipo
- **team_name:** Nombre del equipo

- **former_team_names:** Nombres anteriores del equipo
- **current:** Indicador de equipo activo
- **former:** Indicador de equipo anterior
- **defunct:** Indicador de equipo desaparecido
- **first_appearance:** Año de la primera aparición del equipo

Teams Geo

Descripción: Contiene información geográfica de los equipos, incluyendo ciudad, estadio y coordenadas.

Dimensiones:

- Filas: **143**
- Columnas: **7**

Columnas:

- **team_id:** Identificador del equipo
- **team_name:** Nombre del equipo
- **city:** Ciudad del equipo
- **stadium:** Nombre del estadio
- **latitude:** Latitud del estadio
- **longitude:** Longitud del estadio
- **incompleta:** Indicador de datos incompletos

Los campos de tipo object indican que contienen texto.

3.5. Control de valores nulos

Appearances

- **Columnas con valores nulos:**
 - **subdivision:** 349.980 registros
- **Descripción:** La columna subdivision contiene un alto número de valores nulos, lo que indica que para muchos registros no se especifica una subdivisión dentro de la división. El resto de las columnas se encuentra completo, lo que permite un análisis confiable sobre resultados y estadísticas de los partidos.

Matches

- **Columnas con valores nulos:**
 - **subdivision:** 174.990 registros
- **Descripción:** La columna subdivision presenta valores nulos en aproximadamente la mitad de los registros, reflejando que no siempre se registra la subdivisión en los

partidos. Los demás campos cuentan con información completa, permitiendo un análisis de resultados y márgenes de goles sin inconvenientes.

Seasons

- **Columnas con valores nulos:**
 - **subdivision:** 371 registros
- **Descripción:** Solo la columna subdivision presenta algunos valores nulos, lo que sugiere que en pocas temporadas no se registró la subdivisión correspondiente. El resto de la información está completa, lo que facilita el análisis de ganadores y cantidad de equipos por temporada.

Standings

- **Columnas con valores nulos:**
 - **subdivision:** 8.208 registros
- **Descripción:** La columna subdivision tiene valores nulos en un número reducido de registros, mientras que los datos de posiciones, puntos y resultados están completos, permitiendo un análisis fiable de la clasificación de los equipos.

Teams

- **Columnas con valores nulos:**
 - **former_team_names:** 126 registros
- **Descripción:** La columna former_team_names presenta valores nulos para algunos equipos que no tuvieron nombres anteriores. Esto no afecta la información principal sobre la identidad y estado actual de los equipos.

Teams Geo

- **Columnas con valores nulos:**
 - **latitude:** 10 registros
 - **longitude:** 10 registros
 - **incompleta:** 122 registros
- **Descripción:** Las columnas de latitud y longitud tienen pocos valores nulos, indicando que casi todos los equipos cuentan con coordenadas geográficas registradas. La columna incompleta indica registros con información parcial, principalmente para casos específicos donde faltan datos de ubicación o estadio.

3.6. Valores duplicados por columnas

Appearances

- **key_id**: No presenta duplicados, garantizando la unicidad de cada fila.
- **season_id**: Se detectaron 431 valores repetidos, correspondientes a temporadas con múltiples registros. Por ejemplo, S-1888-1 aparece 264 veces.
- **match_id**: Se encontraron 203.956 valores duplicados, ya que cada partido está registrado para ambos equipos.
- **team_id**: Presenta 144 valores repetidos, reflejando que los equipos participan en múltiples partidos a lo largo de las temporadas.

Matches

- **key_id**: Sin duplicados, asegurando unicidad de cada registro.
- **season_id**: 431 valores repetidos, correspondientes a la repetición de temporadas a lo largo de los partidos.
- **match_id**: Sin duplicados, ya que cada partido tiene un identificador único en esta tabla.

Seasons

- **key_id** y **season_id**: Sin duplicados, garantizando que cada temporada esté representada una sola vez.

Standings

- **key_id**: Sin duplicados.
- **season_id**: 431 valores repetidos, correspondientes a múltiples equipos clasificados por temporada.
- **team_id**: 141 valores repetidos, ya que los equipos aparecen en varias temporadas con su posición en la tabla.

Teams

- **key_id** y **team_id**: Sin duplicados, asegurando unicidad de cada equipo.

Teams Geo

- **team_id**: Sin duplicados.

3.7. Valores duplicados por filas

Se realizó un análisis de duplicados considerando todas las columnas de cada fila para identificar registros exactamente iguales. En este caso, no se encontraron filas duplicadas en ninguno de los archivos, lo que garantiza que cada registro es único a nivel de fila completa.

3.8. Modelo de datos normalizado

Para llevar a cabo el análisis, la limpieza y la organización de los datos utilizando **Pandas** y **Apache Spark**, se aplicó un modelo de datos normalizado.

Este enfoque permitió trabajar cada dataset por separado, manteniendo una estructura clara y evitando redundancia entre tablas.

Al normalizar, se organizaron los datos en entidades independientes, cada una con sus atributos propios y con relaciones lógicas entre sí a través de claves como **team_id**, **match_id** o **season_id**.

La normalización facilitó varias tareas del proceso:

- **Limpieza más precisa:** cada archivo se trató de forma individual, permitiendo eliminar columnas irrelevantes, corregir tipos de datos y manejar valores nulos de manera consistente.
- **Estandarización:** se aseguró que los nombres de columnas, formatos numéricos y estructuras fueran coherentes entre todas las tablas.
- **Integración eficiente:** al mantener la información separada por entidad, fue más sencillo combinar datasets cuando era necesario, especialmente para análisis más complejos o cruces entre equipos, temporadas o resultados.
- **Mayor claridad en el análisis:** tanto **Pandas** como **Apache Spark** se beneficiaron de este diseño, ya que permitieron trabajar con tablas limpias, ordenadas y con un esquema bien definido.

3.9. Limpieza de los archivos

Los archivos fueron sometidos a un proceso de limpieza y transformación con el objetivo de estandarizar y preparar los datos para su análisis.

Matches

1. **Eliminación de columnas irrelevantes:** Se eliminaron columnas que no aportaban valor al análisis, tales como `key_id`, `tier`, `subdivision`, `score`, `home_team_win`, `away_team_win` y `draw`. Esto permitió reducir la complejidad del dataset y enfocarse en las variables más relevantes para el análisis de partidos y resultados.
2. **Conversión de tipos de datos:**
 - a. Columnas numéricas (`season`, `home_team_score`, `away_team_score`, `home_team_score_margin`, `away_team_score_margin`) se convirtieron a tipo **Int64**, permitiendo cálculos y agregaciones posteriores sin pérdida de información.
 - b. Columnas que contienen identificadores o categorías (`season_id`, `division`, `match_id`, `match_name`, `home_team_id`, `home_team_name`, `away_team_id`, `away_team_name`, `result`) se convirtieron a **string**, asegurando consistencia en el manejo de valores categóricos y texto.
3. **Guardado del dataset limpio:** El archivo procesado se almacenó en la carpeta **Refined** como `matches_clean.csv`, asegurando la persistencia de los datos limpios para futuros análisis.
4. **Organización de archivos:** El archivo original fue movido de la carpeta **Landing** a **Raw**, siguiendo la estructura de almacenamiento de datos que separa los datos originales de los refinados.

Resultado final:

- Filas: **203.956**
- Columnas: **14**
- Tipos de datos: 5 columnas **Int64**, 9 columnas **string**
- Memoria utilizada: 22.8 MB
- No se detectaron valores nulos en ninguna columna tras la limpieza, garantizando integridad de los datos.
- Columnas finales:
 - **season_id**
 - **season**
 - **division**
 - **match_id**
 - **match_name**
 - **home_team_id**
 - **home_team_name**
 - **away_team_id**
 - **away_team_name**
 - **home_team_score**
 - **away_team_score**
 - **home_team_score_margin**
 - **away_team_score_margin**

- **result**

Seasons

1. **Eliminación de columnas irrelevantes:** Se eliminaron las columnas **key_id**, **tier** y **subdivision**, que no aportaban información relevante para el análisis de las temporadas, simplificando la estructura del dataset.
2. **Conversión de tipos de datos:**
 - a. Columnas numéricas (**season**, **count_teams**) se convirtieron a tipo **Int64**, permitiendo realizar cálculos y agregaciones de manera precisa.
 - b. Columnas categóricas o de texto (**season_id**, **division**, **winner**) se convirtieron a **string**, asegurando consistencia en el manejo de identificadores y nombres.
3. **Guardado del dataset limpio:** El archivo procesado se almacenó en la carpeta **Refined** como **seasons_clean.csv**, garantizando la persistencia de los datos limpios para análisis posteriores.
4. **Organización de archivos:** El archivo original fue movido de la carpeta **Landing** a **Raw**, siguiendo la estructura de almacenamiento que separa los datos originales de los refinados.

Resultado final:

- Filas: **431**
- Columnas: **5**
- Tipos de datos: 2 columnas **Int64**, 3 columnas **string**
- Memoria utilizada: 17.8 KB
- No se detectaron valores nulos en ninguna columna tras la limpieza, asegurando la integridad de los datos.
- Columnas finales:
 - **season_id**
 - **season**
 - **division**
 - **winner**
 - **count_teams**

Teams

1. **Eliminación de columnas irrelevantes:** Se eliminaron las columnas **key_id** y **subdivision**, que no aportaban información útil para el análisis de los equipos, simplificando la estructura del dataset.
2. **Conversión de tipos de datos:**

- o Columnas numéricas (current, former, defunct, first_appearance) se convirtieron a tipo **Int64**, lo que permite realizar cálculos y agregaciones de manera consistente.
 - o Columnas de texto o categóricas (team_id, team_name, former_team_names) se convirtieron a **string**, asegurando uniformidad en el manejo de identificadores y nombres de equipos.
3. **Guardado del dataset limpio:** El archivo procesado se almacenó en la carpeta **Refined** como teams_clean.csv, asegurando la persistencia de los datos limpios para análisis posteriores.
 4. **Organización de archivos:** El archivo original fue movido de la carpeta **Landing** a **Raw**, respetando la estructura de almacenamiento que separa los datos originales de los refinados.

Resultado final:

- Filas: **144**
- Columnas: **7**
- Tipos de datos: 4 columnas **Int64**, 3 columnas **string**
- Memoria utilizada: 8.6 KB
- Valores nulos: la columna former_team_names contiene 126 valores nulos, correspondientes a equipos que no tuvieron nombres anteriores. El resto de las columnas se encuentra completo.
- Columnas finales:
 - o **team_id**
 - o **team_name**
 - o **former_team_names**
 - o **current**
 - o **former**
 - o **defunct**
 - o **first_appearance**

Standings

1. **Eliminación de columnas irrelevantes:** Se eliminaron las columnas key_id, tier y subdivision, que no aportaban información útil para el análisis de las posiciones de los equipos, simplificando la estructura del dataset.
2. **Conversión de tipos de datos:**
 - o Columnas numéricas (season, position, played, wins, draws, losses, goals_for, goals_against, goal_difference, points, point_adjustment) se convirtieron a tipo **Int64**, permitiendo realizar cálculos y agregaciones consistentes sobre resultados y estadísticas.

- Columnas de texto o categóricas (season_id, division, team_id, team_name) se convirtieron a **string**, asegurando uniformidad en el manejo de identificadores y nombres de equipos.
- 3. **Guardado del dataset limpio:** El archivo procesado se almacenó en la carpeta **Refined** como standings_clean.csv, garantizando la persistencia de los datos limpios para análisis posteriores.
- 4. **Organización de archivos:** El archivo original fue movido de la carpeta **Landing** a **Raw**, siguiendo la estructura de almacenamiento que separa los datos originales de los refinados.

Resultado final:

- Filas: **9.556**
- Columnas: **15**
- Tipos de datos: 11 columnas **Int64**, 4 columnas **string**
- Memoria utilizada: 1.2 MB
- Valores nulos: No se detectaron valores nulos en ninguna columna tras la limpieza, asegurando integridad de los datos.
- Columnas finales:
 - **season_id**
 - **season**
 - **division**
 - **position**
 - **team_id**
 - **team_name**
 - **played**
 - **wins**
 - **draws**
 - **losses**
 - **goals_for**
 - **goals_against**
 - **goal_difference**
 - **points**
 - **point_adjustment**

Teams Geo

1. **Eliminación de columnas irrelevantes:** Se eliminó la columna incompleta, que indicaba registros con información parcial, ya que no aportaba valor directo al análisis geográfico.
2. **Normalización de valores faltantes:** Se reemplazaron valores inconsistentes o marcadores de datos faltantes como "?", "N/A", "-", "null" y "<NA>" por **NA** de Pandas.

La columna stadium fue completada con "Unknown" para registros vacíos, asegurando consistencia en los nombres de estadios.

3. Conversión de tipos de datos:

- o Columnas de coordenadas (latitude, longitude) se convirtieron a **Float64**, permitiendo cálculos geográficos y visualizaciones en mapas.
- o Columnas de texto o categóricas (team_id, team_name, city, stadium) se convirtieron a **string**, garantizando consistencia en identificadores y nombres.

4. Guardado del dataset limpio:

El archivo procesado se almacenó en la carpeta **Refined** como **teams_geo_clean.csv**, asegurando que los datos limpios estuvieran disponibles para análisis geoespaciales.

5. Organización de archivos:

El archivo original fue movido de la carpeta **Landing** a **Raw**, manteniendo la estructura de almacenamiento que separa los datos originales de los refinados.

Resultado final:

- Filas: **143**
- Columnas: **6**
- Tipos de datos: 2 columnas **Float64**, 4 columnas **string**
- Memoria utilizada: 7.1 KB
- Valores nulos: las columnas latitude y longitude presentan 25 y 24 valores nulos respectivamente, correspondientes a registros de equipos con coordenadas faltantes.
- Columnas finales:
 - o **team_id**
 - o **team_name**
 - o **city**
 - o **stadium**
 - o **latitude**
 - o **longitude**

Evolution Points

1. **Creación del dataset:** Se generó un nuevo archivo basado en el dataset de **appearances**, registrando la información de cada temporada y la evolución de los puntos asignados según el resultado de los partidos **win, draw o lose**.
2. Conversión de tipos de datos:
 - o La columna season_id se convirtió a **string** para mantener consistencia en los identificadores de temporada.
 - o La columna points se convirtió a **Int64**, permitiendo cálculos y agregaciones.
 - o La columna result se convirtió a **string**, representando los distintos resultados de los partidos **win, draw o lose**.

3. **Guardado del dataset limpio:** El archivo procesado se almacenó en la carpeta Refined como **evolution_points.csv**, asegurando que los datos estuvieran disponibles para análisis posteriores.
4. **Organización de archivos:** El archivo original (**appearances**) fue movido de la carpeta **Landing** a **Raw**, manteniendo la estructura de almacenamiento que separa los datos originales de los refinados.

Resultado final:

- Filas: **1293**
- Columnas: **3**
- Tipos de datos: 1 columna **Int64**, 2 columnas **string**
- Memoria utilizada: 40.4 KB
- Valores nulos: No se detectaron valores nulos en ninguna columna tras la limpieza, asegurando integridad de los datos.
- Columnas finales:
 - **season_id**
 - **points**
 - **result**

Appearances

El archivo **Appearances** fue descartado del análisis ya que la información que contenía podía ser obtenida principalmente del archivo **Matches**, que incluye todos los detalles relevantes sobre partidos, equipos, resultados y estadísticas asociadas.

Dado que **Matches** proporciona un registro completo de los encuentros y permite calcular métricas equivalentes a las presentes en **Appearances**, se decidió prescindir de este archivo para simplificar el flujo de datos y evitar redundancia en el procesamiento y almacenamiento. Esta decisión asegura que el análisis se centre en los datasets más completos y consistentes, optimizando el uso de recursos y facilitando la integración de la información.

3.10. Resumen de limpieza y preparación de datos

El proceso de limpieza y normalización garantiza que los datasets Matches, Seasons, Teams, Standings y Teams Geo se encuentren normalizados, consistentes y completos, listos para su integración y para análisis avanzados.

Asimismo, los datos geográficos de Teams Geo han sido preparados para visualizaciones espaciales y análisis geoespaciales precisos de los equipos, asegurando la calidad y confiabilidad de la información utilizada en todo el proyecto.

El archivo Appearances fue descartado, dado que la información que contenía podía obtenerse principalmente desde Matches, evitando redundancia y optimizando el flujo de datos.

4. Estructura del modelo para Hive

Una vez finalizada la fase de limpieza y preparación de datos utilizando **Apache Spark**, los datasets resultantes quedaron listos para ser modelados y organizados dentro de un entorno como **Hive**.

Dado que todos los archivos ya están almacenados en **HDFS**, se optó por crear tablas externas para cada uno de los datasets. Este enfoque permite que **Hive** lea y consulte los datos directamente desde su ubicación en HDFS, pero sin modificarlos ni eliminarlos.

Tablas que se crearían:

- **Matches**: información completa de los partidos (equipos, resultados, goles, temporada)
- **Seasons**: datos de cada temporada (división, cantidad de equipos, ganadores).
- **Teams**: información de equipos, sus nombres, estado, fecha de aparición.
- **Standings**: rendimiento de los equipos por temporada (puntos, victorias, derrotas, goles, posición).
- **Teams_geo**: información geográfica de cada equipo (ciudad, país, estadio, coordenadas).

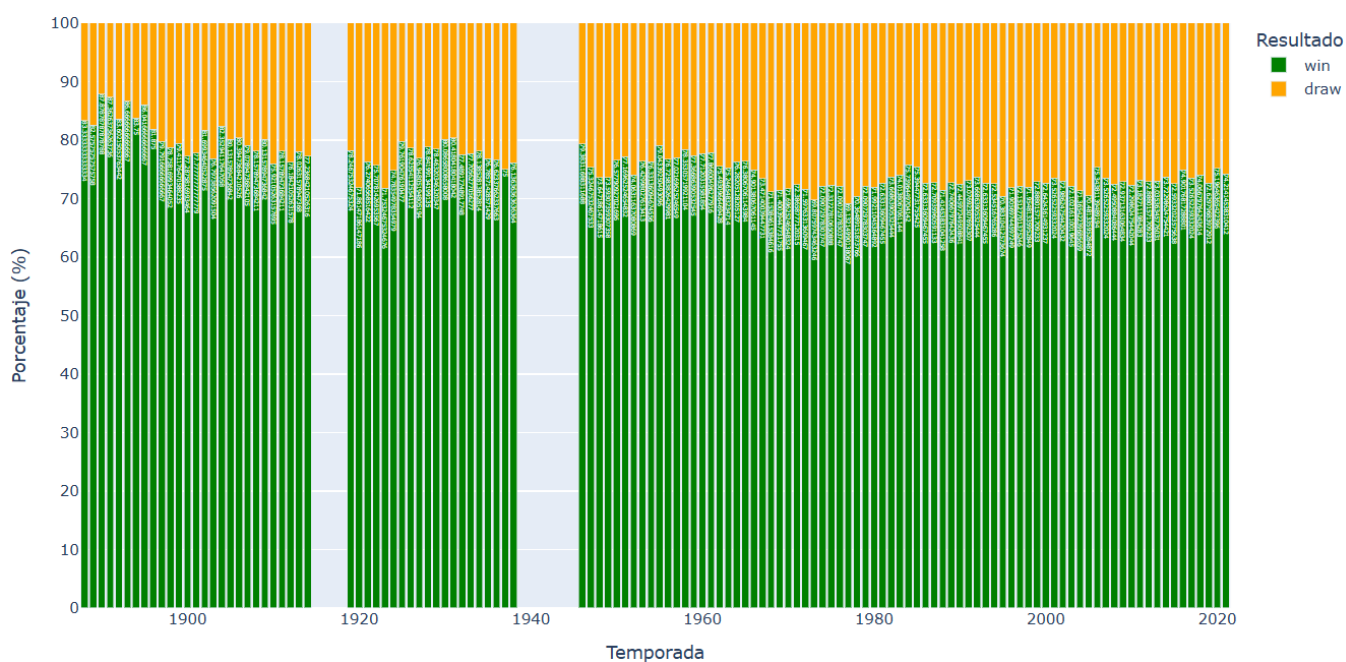
5. Desarrollo de análisis

De acuerdo con los datos proporcionados y datos externos agregados, nos planteamos responder las siguientes preguntas:

1. Analizar la evolución de la cantidad de partidos por temporada.

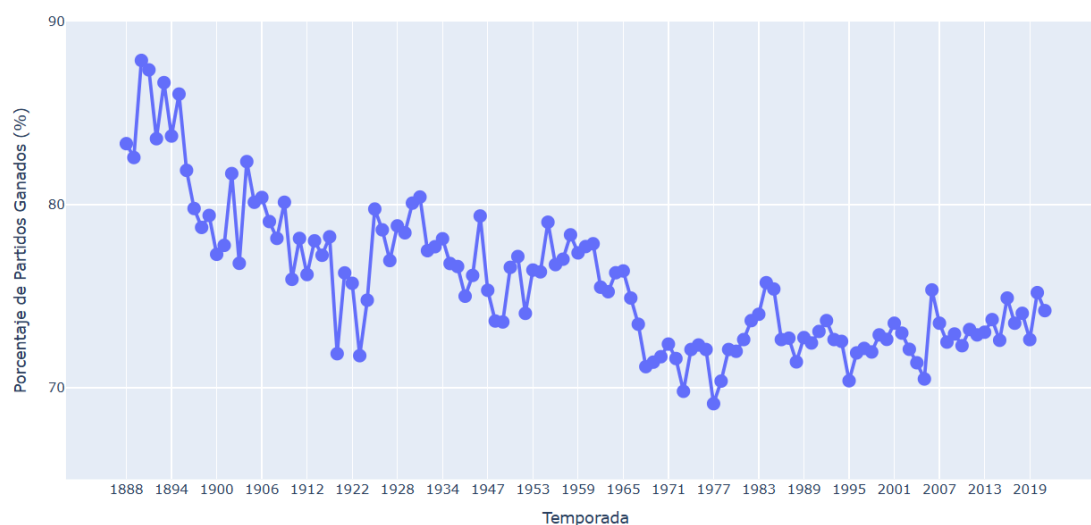
1.1. Visualizaciones

1.1.1. Evolución en el tiempo del porcentaje de partidos, según finalizaron con un empate o victoria.



1.1.2. Evolución del porcentaje de partidos que finalizaron con victoria.

Evolución del porcentaje de partidos ganados por temporada



1.2. Análisis descriptivo de los datos

A partir de esta primera apertura (grafico 1.1.1) se observan dos aspectos:

- Se observa que antes del 1900 el porcentaje de partidos con victorias en cada temporada superaba el 82% del total, desde 1930 este porcentaje no supera el 80%.
- Para las temporadas del 1915 al 1918 y 1939 al 1945 no se jugaron debido a los conflictos bélicos por la primera y segunda guerra mundial.

En cuanto al primer punto del mayor porcentaje de partidos con victorias antes del 1.900 entendimos ampliar la información (grafico 1.1.2):

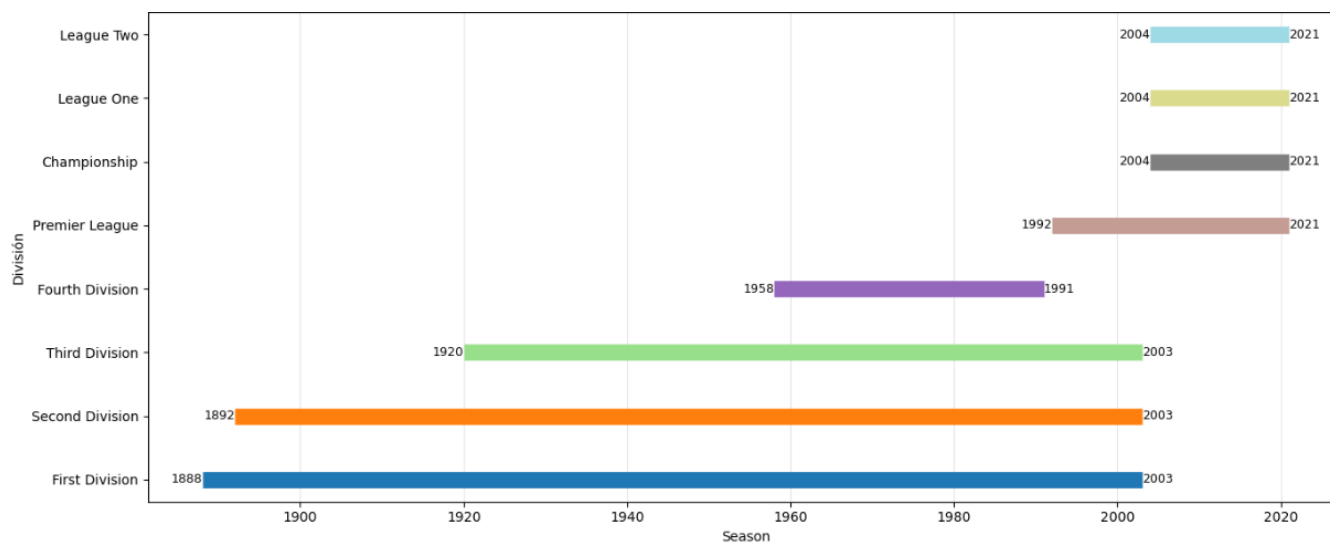
Para justificar esta tendencia estuvimos indagando sobre los motivos, e identificamos que, en los inicios del fútbol inglés, la principal razón para los pocos empates era la inexistencia de una regla universalmente aceptada para el desempate. En lugar de un empate, los partidos a menudo se jugaban hasta que un equipo marcaba un gol más, o en algunos casos, se recurría a sorteos o partidos de desempate, aunque estas prácticas variaron.

- **Falta de estandarización:** Al principio, las reglas variaban entre los diferentes clubes y escuelas, lo que llevaba a diferentes procedimientos para resolver los partidos.
- **Poca importancia del empate:** El empate no era el resultado buscado, ya que a los equipos les interesaba más una victoria clara y una forma definitiva de pasar a la siguiente ronda en torneos.
- **Prácticas de desempate:** Para desempatar, a veces se jugaba de nuevo el partido o se realizaban sorteos. Estas prácticas, aunque diferentes a la regla actual, no resultaban en el empate de forma automática.

2. Examinar la evolución histórica de las distintas categorías del fútbol inglés.

2.1. Visualizaciones

2.1.1. Diagrama de Gantt de participación por división



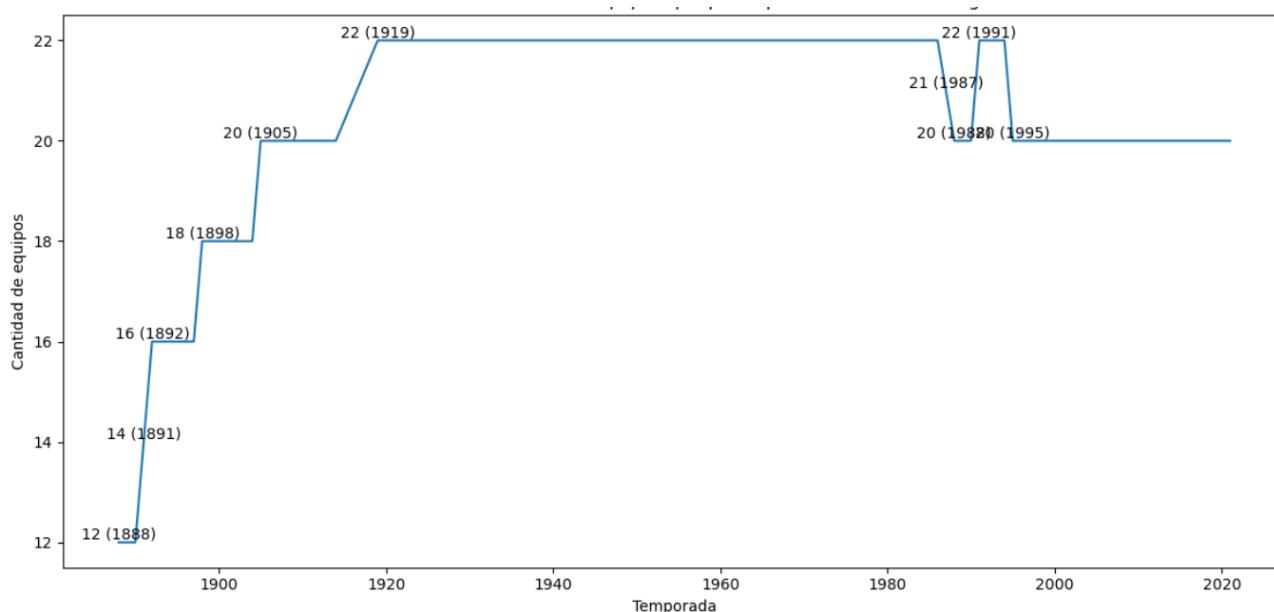
2.2. Análisis descriptivo de datos

A lo largo del tiempo, la estructura de las divisiones del fútbol inglés ha experimentado varias modificaciones. El cambio más significativo ocurrió en 1992 con la creación de la Premier League, que sustituyó a la First Division como la máxima categoría. Esta última continuó funcionando como segunda división hasta 2003, cuando fue reemplazada en 2004 por la Championship.

3. Estudiar la variación en la cantidad de equipos participantes en la máxima categoría.

3.1. Visualizaciones

3.1.1. Evolución de cantidad de equipos que participan en la First Division según temporada



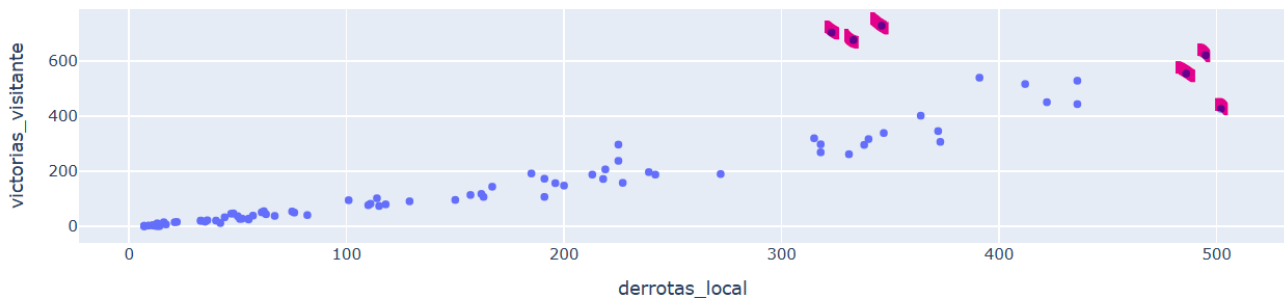
3.2. Análisis descriptivo de datos

El gráfico muestra cómo la cantidad de equipos en la en la máxima categoría fue aumentando en los primeros años, acompañando la expansión inicial del fútbol organizado. Luego la cifra se estabiliza durante un largo período, reflejando una estructura de liga prácticamente fija. Hacia las últimas temporadas aparecen variaciones más marcadas, señalando procesos de reestructuración y cambios en el formato de la competencia.

4. Explorar la distribución de equipos según sus victorias como visitantes y derrotas como locales.

4.1. Visualizaciones

4.1.1. Distribución de equipos según cantidad de derrotas como local vs victorias en la First División



4.2. Análisis descriptivo de datos

El gráfico muestra la relación entre las derrotas como local y las victorias como visitante de los equipos de la máxima categoría. Se observa una tendencia creciente: los equipos que acumulan más derrotas en casa también suelen registrar más triunfos fuera de ella, lo que refleja principalmente una trayectoria histórica más extensa y un mayor número total de partidos disputados.

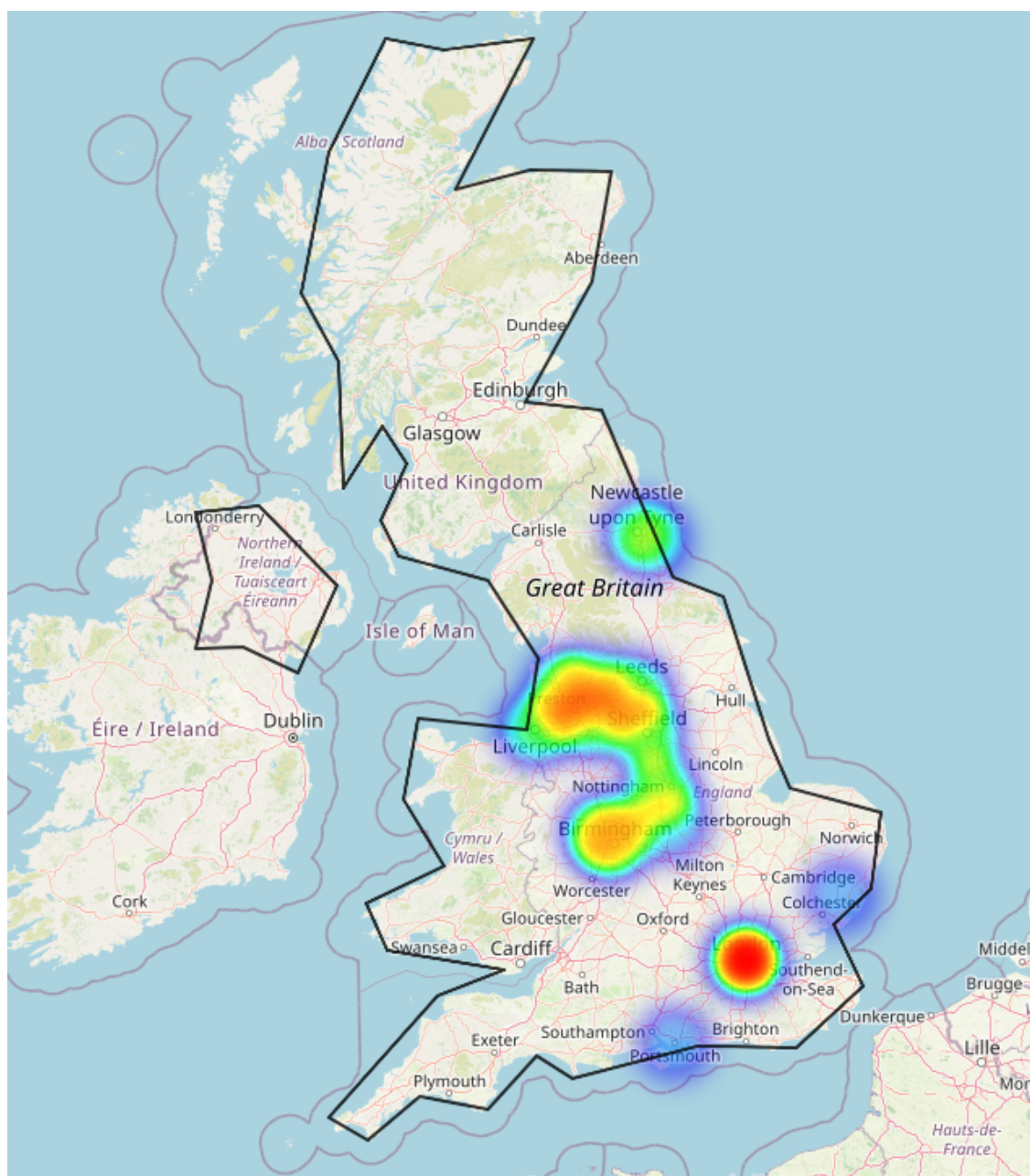
Los puntos destacados en color señalan equipos con una participación excepcionalmente alta en la liga, lo que los convierte en **outliers** dentro de la distribución. De izquierda a derecha, estos casos corresponden a **Manchester United, Arsenal, Liverpool, Aston Villa, Everton y West Bromwich Albion**, clubes con una larga permanencia en la categoría y un volumen de encuentros significativamente superior al promedio.

Para este punto creemos importante señalar la situación del West Bromwich Albión un club tradicional que en la actualidad sufre problemas económicos, pero fue campeón en la temporada 1919-20 de la First Division, y para las temporadas del 1953-54 sus rendimientos lo llevo a que los medios propusieran a ese equipo para representar a la selección inglesa.

5. Determinar geográficamente la concentración de los clubes campeones de la máxima categoría.

5.1. Visualizaciones

5.1.1. Mapa de calor según las ciudades con equipos que tienen equipos ganadores de la máxima categoría



5.2. Análisis descriptivo de datos

El mapa de calor muestra que los campeones del fútbol inglés se concentran en el centro-norte del país, especialmente en ciudades industriales como Manchester, Liverpool, Leeds, Nottingham y Birmingham. Estas regiones fueron históricamente el corazón del fútbol inglés, donde surgieron clubes fuertes ligados a centros urbanos industriales con grandes comunidades obreras.

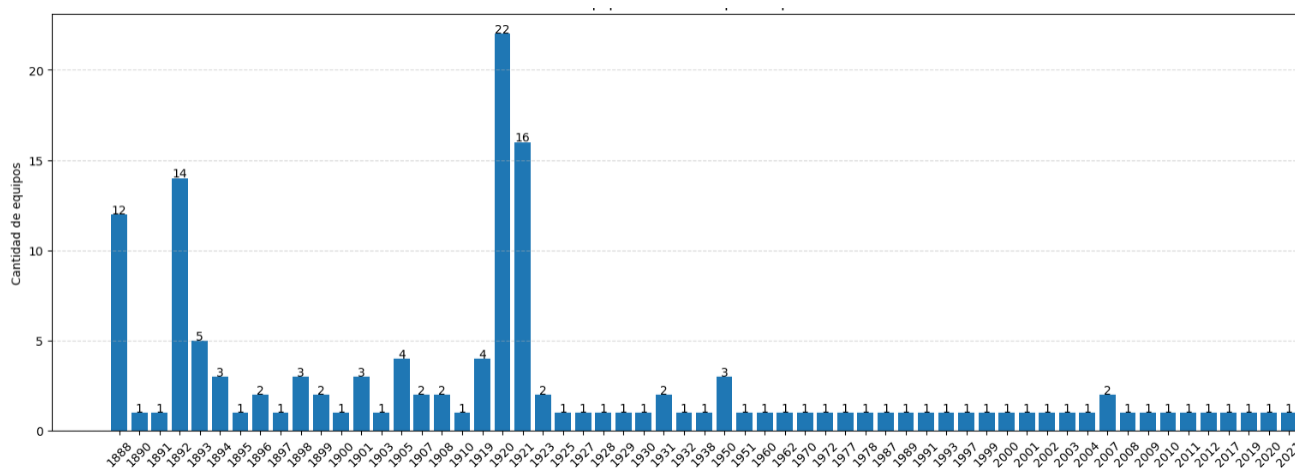
En contraste, el sur de Inglaterra (excepto Londres) que surgen con Chelsea, Arsenal, y Tottenham como excepciones. En conjunto, el mapa evidencia que la distribución de campeones no es aleatoria, sino que responde a patrones sociales e industriales del desarrollo de Inglaterra.

Para una mejor visualización del mapa puede encontrarlo aquí: **...\229705-259620\Notebooks\Pandas\mapa_ciudades_inglaterra_HeatMap.html**

6. Analizar la frecuencia y el ritmo de crecimiento en la cantidad de equipos que han participado en las diferentes divisiones del fútbol inglés

6.1. Visualizaciones

6.1.1. Apertura de equipos debutantes por temporada



6.2. Análisis descriptivo de los datos

Los picos que aparecen en las temporadas 1888, 1892, 1920 y 1921 no son errores. Cada uno corresponde a momentos clave de expansión en la historia del fútbol inglés, cuando la liga incorporó muchos clubes simultáneamente.

Temporada 1888 (12 clubes debutan)

Marca la fundación de la Football League, la primera liga profesional del mundo. Los 12 clubes fundadores comenzaron a competir ese mismo año, por lo que todos figuran como debutantes juntos.

Temporada 1892 (14 clubes debutan)

Ocurre una gran ampliación del sistema, vinculada a la creación de una nueva división. Esto permitió el ingreso de 14 equipos nuevos en una sola temporada.

Temporada 1920 (22 clubes debutan)

La Football League absorbió casi por completo una liga regional muy fuerte, incorporando de una vez a gran parte de sus equipos. Como resultado, 22 clubes debutaron simultáneamente.

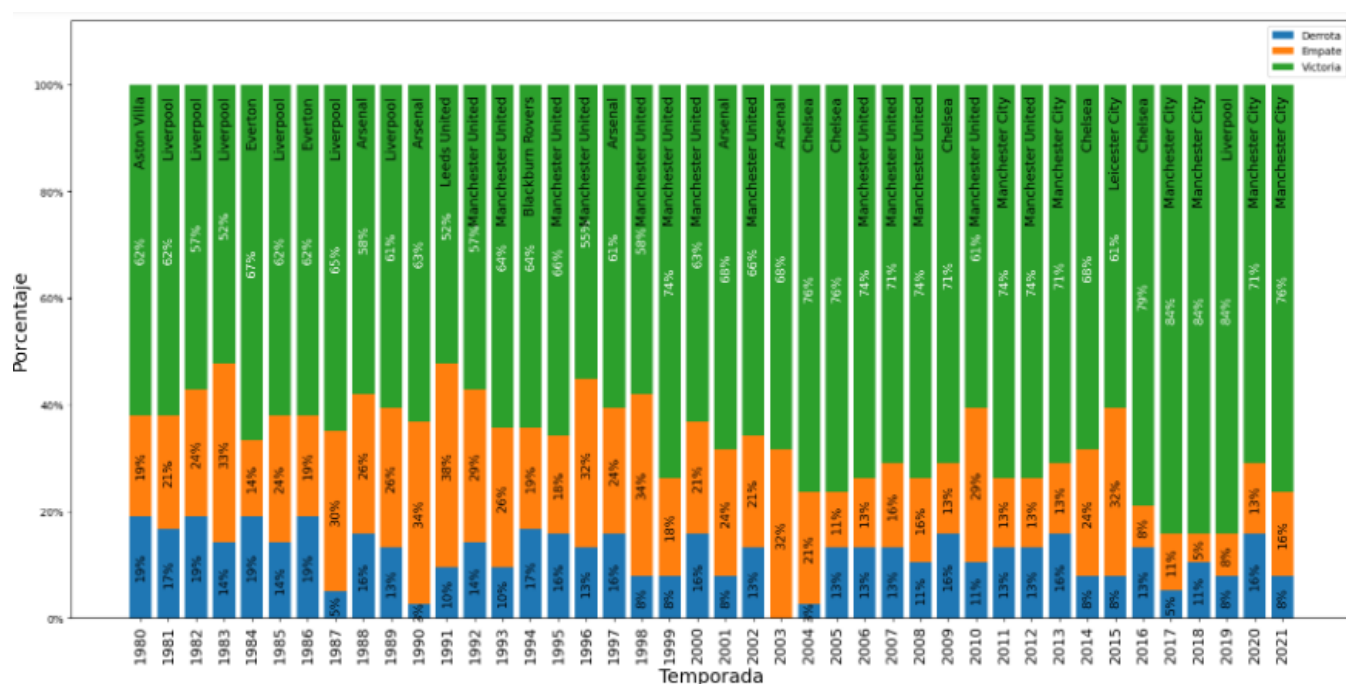
Temporada 1921 (16 clubes debutan)

Se creó una división adicional para equilibrar geográficamente el sistema. Esto generó otro ingreso masivo.

7. Calcular el porcentaje promedio de victorias de los clubes campeones de la máxima categoría desde 1980.

7.1. Visualizaciones

7.1.1. Porcentaje de victorias de los clubes campeones en cada temporada desde 1980



7.2. Análisis descriptivo de los datos

De acuerdo con el resultado del cruce de información se puede notar que en promedio el porcentaje de victorias de los clubes campeones en cada temporada desde 1980 es un 67%.

Se pueden destacar 2 puntos:

1. El alto porcentaje de victorias para el Manchester City en las temporadas 2017 y 2018 y para el Liverpool en 2019, ambos con un 85%.

El motivo por el cual los llevo a esto es que ambos equipos estaban en la cima de la Premier en disputa del título, con estilos muy ofensivos en la forma de jugar, para el Manchester City su DT era Guardiola y Liverpool Klopp, además que equipos como Arsenal, Chelsea, Tottenham y Manchester United estaban más irregulares en esas temporadas.

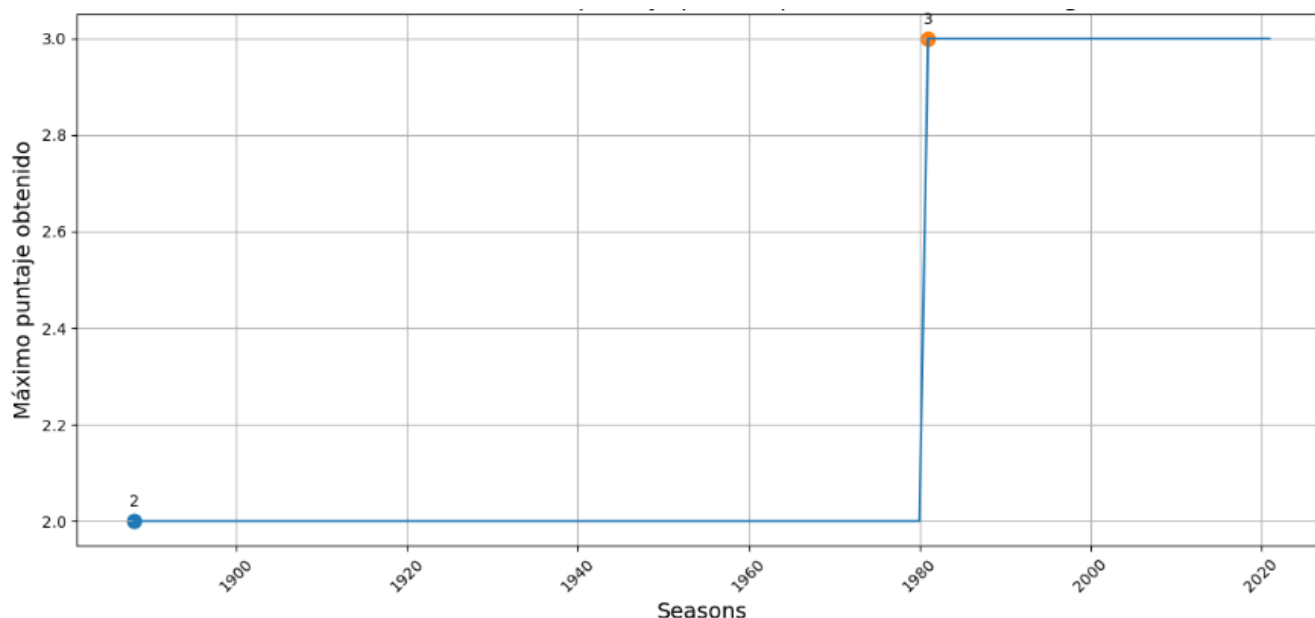
Para el caso de Manchester City en la temporada 2017 batió el récord de la cantidad de puntos realizada con 100 puntos y Liverpool obtuvo un récord en 2019 ganando 26 de sus primeros 27 partidos.

2. En 2003 se identificó que el Arsenal finalizó la temporada sin derrotas, analizando el comportamiento histórico, se observa que, esta situación no se daba desde 1.888 en el primer campeonato de Inglaterra, por tal motivo es importante destacarlo.

8. En que temporada se comienza a conceder 3 puntos por partidos ganados.

8.1. Visualizaciones

8.1.1. Evolución histórica del puntaje máximo obtenido por partido del futbol ingles



8.2. Análisis descriptivo de datos

De acuerdo con el cruce de información realizado, se observa que entre 1888 y 1980 por cada victoria se otorgaban 2 puntos, mientras que a partir de la temporada 1981 se comenzó a conceder 3 puntos por partido ganado. Al profundizar en el análisis, identificamos que este cambio respondió a la intención de promover un fútbol más ofensivo. La FIFA impulsó esta modificación en 1980 luego de evaluar que en los torneos internacionales, especialmente en los mundiales, se registraba una baja cantidad de goles. En consecuencia, el fútbol inglés adoptó la regla a partir de 1981, tal como se refleja en los datos analizados, y hacia 1990 la medida ya se había estandarizado a nivel global.

Un elemento adicional para destacar es que, si observamos el gráfico 1.1.2, se aprecia que en la liga inglesa entre las temporadas 1965 y 1980 el porcentaje de partidos ganados era relativamente bajo. Este comportamiento podría interpretarse como uno de los motivos que llevaron a introducir el sistema de 3 puntos, buscando incentivar a los equipos a adoptar un estilo de juego más ofensivo y competitivo.

7. Dashboard en Power BI

Como parte del proceso de análisis y visualización de los resultados obtenidos, se utilizó **Power BI** para la construcción de un dashboard que permitiera comunicar de forma efectiva los hallazgos más relevantes.

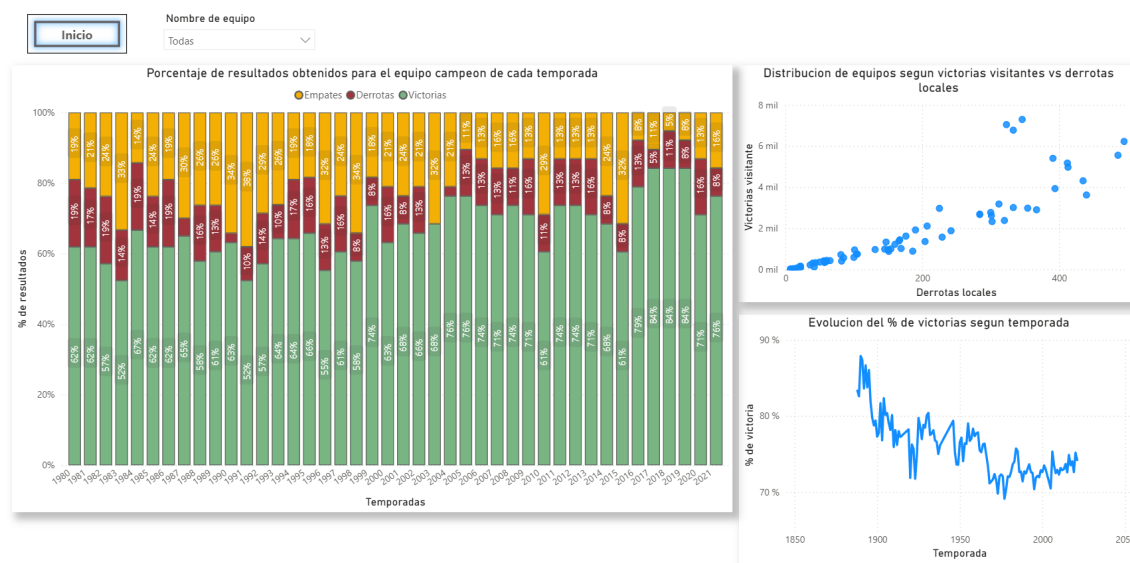
Para ello, se seleccionaron tres preguntas clave del desarrollo analítico, priorizando aquellas que ofrecían una perspectiva representativa sobre la evolución competitiva y el comportamiento histórico del fútbol inglés.

Las preguntas elegidas fueron:

1. Analizar la evolución de la cantidad de partidos por temporada, desglosando los resultados en empates y victorias, con el fin de identificar patrones competitivos a lo largo del tiempo.
2. Calcular el porcentaje promedio de victorias de los clubes campeones de la máxima categoría desde 1980, para evaluar el nivel de dominio competitivo requerido para obtener el título en distintos períodos históricos.
3. Explorar la distribución de equipos según sus victorias como visitantes y derrotas como locales, para identificar comportamientos competitivos relevantes.

Para la construcción de las visualizaciones correspondientes, se generaron tres archivos CSV derivados del procesamiento previo de los datos: **porcentaje_victorias_por_temporadas**, **derrotas_locales_vs_victorias_visitantes** y **df_summary**, los cuales sirvieron como fuentes estructuradas para alimentar el dashboard y posibilitar una interpretación clara y fundamentada de los resultados.

Acceso: [Dashboard Power BI](#)



8. Conclusiones

El desarrollo de este proyecto permitió una exploración profunda y estructurada de los datos del fútbol inglés, aplicando un flujo completo que incluyó limpieza, transformación, modelado y análisis. El trabajo con **Pandas** y luego la replicación del análisis con **Apache Spark** resultó especialmente enriquecedor, ya que permitió comparar directamente las fortalezas y limitaciones de ambas herramientas en distintos contextos.

Pandas se mostró ideal para la etapa inicial de exploración y transformación de los datos, facilitando tareas como filtrado de registros, agrupamientos, manejo de valores nulos y conversión de tipos de datos. Su facilidad de uso y rapidez para trabajar localmente con datasets de tamaño moderado permitió comprender a fondo la estructura de la información y sentar las bases para análisis posteriores.

Apache Spark requirió un mayor nivel de preparación y conocimientos vinculados a entornos distribuidos, pero demostró un gran potencial al procesar grandes volúmenes de datos de manera eficiente y escalable, lo cual es fundamental para proyectos que requieren manejo de datasets de mayor magnitud o integración de múltiples fuentes.

Un aspecto central del trabajo fue el modelado de los datos utilizando un enfoque normalizado, que permitió organizar la información en tablas independientes y consistentes, como **matches**, **seasons**, **teams**, **standings** y **teams_geo**. Esta estructura no solo facilitó la limpieza y la transformación de los datos, sino que también hizo más claro y eficiente el análisis posterior, permitiendo relacionar las entidades de manera lógica mediante claves.

El modelado reforzó la importancia de pensar cuidadosamente la estructura de los datos antes de comenzar a analizar, ya que un diseño adecuado garantiza resultados consistentes y facilita futuras consultas o integraciones.

Finalmente, la integración de visualizaciones interactivas en Power BI complementó el análisis técnico, permitiendo interpretar los resultados de forma más intuitiva.

En conjunto, este proyecto no solo permitió obtener conclusiones relevantes sobre los datos del fútbol inglés, sino que también ofreció una experiencia práctica muy valiosa para comprender cuándo y cómo utilizar **Pandas** o **Apache Spark** según el tamaño y la complejidad de los datasets, la importancia del modelado normalizado y el valor de complementar el análisis con visualizaciones interactivas. Además, dejó aprendizajes claros sobre la planificación de preguntas analíticas y la organización de datos para proyectos futuros de mayor escala y complejidad.

9. Uso de inteligencia artificial

Durante el desarrollo de este trabajo, se utilizaron herramientas de **inteligencia artificial** para optimizar algunas etapas del proceso, mejorando la eficiencia y la claridad en la elaboración del documento.

Entre los principales usos se destacan:

- **Redacción de textos:** Se empleó para perfeccionar la redacción de distintas secciones del documento, tales como el análisis de datos, la carga de información, la descripción de las columnas de cada archivo, el desarrollo de análisis y las conclusiones obtenidas.
- **Formulación de comparaciones:** Se consultó sobre cómo contrastar los resultados obtenidos utilizando **Pandas** y **Apache Spark**, asegurando que los textos reflejaran de manera precisa las similitudes y diferencias entre ambas herramientas.
- **Resolución de dudas técnicas:** Se solicitó orientación sobre conceptos relacionados con **Big Data**, incluyendo el modelado de datos, la preparación de la información para análisis y visualizaciones, y la realización de consultas específicas en **Pandas** o **Apache Spark** para generar visualizaciones relevantes.

Es importante destacar que, aunque se utilizó inteligencia artificial como apoyo, todas las decisiones finales sobre el contenido del documento, las consultas realizadas y la interpretación de los resultados fueron tomadas por los integrantes del trabajo. Esto garantiza que el documento refleje de manera fiel los aprendizajes y el esfuerzo del equipo.