English Premier League Match Events and Results

Javier De La Fuente - Benjamín Mancilla - Dylan Riquelme
Grupo 12

Descripción:

Corresponde a un dataset en torno al fútbol (para variar) de la liga inglesa, la Premier League. Se presentan datos de los partidos desde la temporada 2001-2002 hasta aproximadamente la temporada 2021-2022, además de los *aggregated stats* como estadísticas agregadas sobre asistencias y goleadores.

Nace de la curiosidad del usuario Joseph Mohr por ver si existen ciertos patrones dentro de esta liga, por ejemplo, observar si un equipo tiende a obtener más tarjetas enfrentándose a ciertos equipos o notar qué equipos pierden o ganan por la marca más alta de goles.

Link: English Premier League Matches Events and Results

Motivación:

El problema que se intentará resolver es el obtener lo más posible de los datos brindados por Joseph. Hay muchos amantes del fútbol que les encantaría saber lo más posible de su equipo o liga amada.
Así, este proyecto podría funcionar como inspiración para cualquier grupo de datos de distintas ligas de fútbol del mundo y cualquier aficionado podría obtener los datos más interesantes de su jugador, equipo o liga favoritos.

Se eligió la opción del **desarrollo de una aplicación web**.

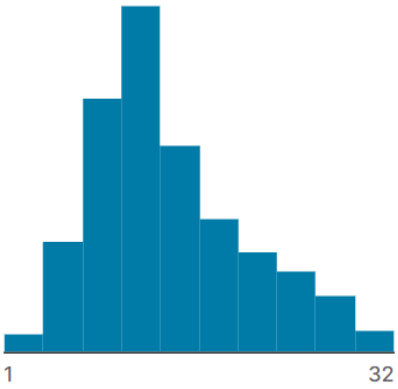
Exploración de Datos:

Tabla	Tuplas totales
all_tables.csv	420
events.csv	633.924
matches.csv	7979

Distribución de “W” de all_tables.csv:

W

Games Won

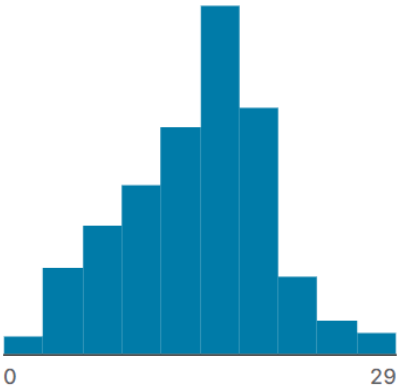


Valid	420	100%
Mismatched	0	0%
Missing	0	0%
Mean	14.3	
Std. Deviation	6.01	
Quantiles	1	Min
	10	25%
	13	50%
	18	75%
	32	Max

Distribución de “L” de all_tables.csv:

L

Games Lost

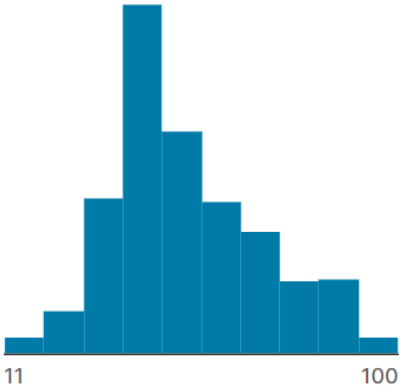


Valid	420	100%
Mismatched	0	0%
Missing	0	0%
Mean	14.3	
Std. Deviation	5.58	
Quantiles	0	Min
	10	25%
	15	50%
	18	75%
	29	Max

Distribución de “Points” de all_tables.csv:

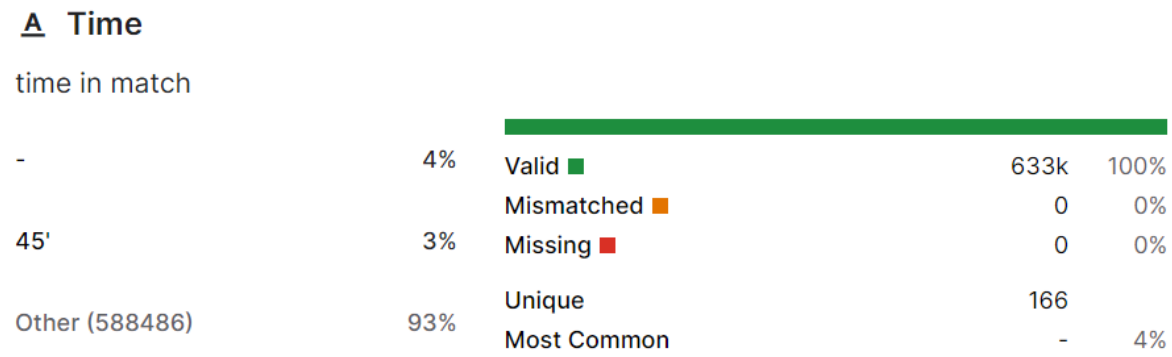
P

Points

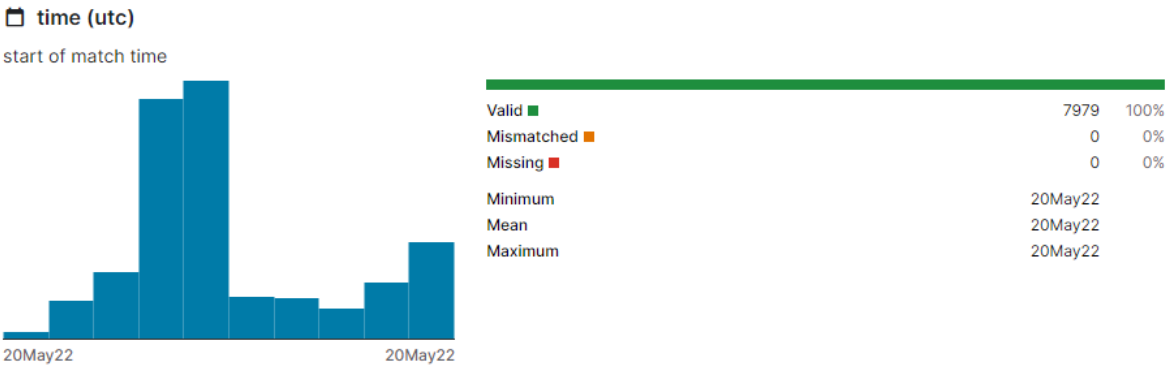


Valid	420	100%
Mismatched	0	0%
Missing	0	0%
Mean	52.2	
Std. Deviation	17.2	
Quantiles	11	Min
	40	25%
	48	50%
	64	75%
	100	Max

Distribución de “Time” (tiempo del evento) de events.csv:



Distribución de “time (utc)” (inicio del encuentro) de matches.csv:



Atributos all_tables.csv	Dominio
Place	Int
Team	String
GP (Games Played)	Int
W	Int
D	Int
L	Int
GF	Int
GA	Int
GD	Int
Points	Int
Year	Int

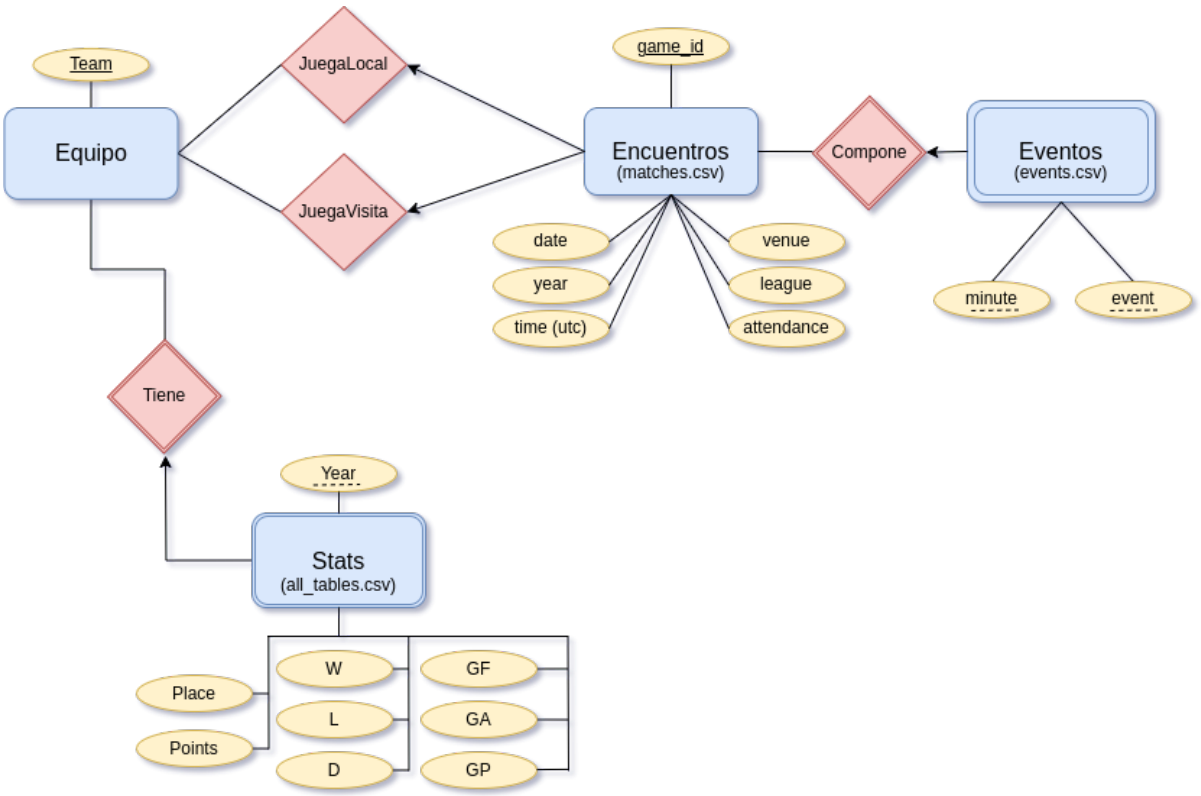
Atributos events.csv	Dominio
id	Int
Time	String
Event	String

Atributos matches.csv (originalmente hay 155 columnas, se tomarán las 9 principales)	Dominio
id	Int
home	String
away	String
date	String
year	Int
time (utc)	String
Attendance	Int
Venue	String
League	String

Consultas (querys) posibles:

- ¿Qué equipo quedó primero en la tabla en la temporada 2009-2010?
- ¿Se producen más eventos en ciertos estadios?
- ¿Cuántas tarjetas rojas se sacaron el 2001 en el estadio Anfield?
- ¿En qué estadio se tiene mayor registro de tarjetas rojas?
- ¿La hora de inicio del partido incide en la cantidad de eventos?
- ¿Cuál fue el partido con más asistencia en 2005?
- ¿Cuántos partidos ganó el Arsenal en un sábado del 2002?

Diagrama E/R:



Traducción a modelo relacional:

Team(team)
Stats(year, Te.team, place, points, w, l, d, gp, gf, ga)
Encuentros(game_id, Te.homeTeam, Te.awayTeam, date, year, time_utc, venue, league, attendance)
Eventos(En.game_id, minute, event)

¿2NF?¿3NF?¿BCNF?:

Ahora, habiendo eliminado la dependencia transitiva de la diferencia de goles, todas las tablas se encuentran en BCNF dado que no hay dependencias funcionales parciales, no hay dependencias transitivas y se cumple que la dependencia funcional viene desde la super llave.

Además, hay que notar que se optó por crear una nueva tabla **Team**, con el objeto de evitar el choque entre la llave primaria de la tabla ahora llamada **Stats** y la tabla **Encuentros**, dadas las relaciones **JuegaLocal** y **JuegaVisita**.

Scripts SQL:

Crear el esquema en el que se trabajará:

```
CREATE SCHEMA fuchibol;
```

Crear las tablas sacadas del modelo relacional:

```
CREATE TABLE fuchibol.Team (
   team VARCHAR (255) PRIMARY KEY
);
```

```
CREATE TABLE fuchibol.Stats (
   year INTEGER,
   team VARCHAR (255) REFERENCES fuchibol.Team(team),
   place SMALLINT,
   points SMALLINT,
   won SMALLINT,
   lost SMALLINT,
   drew SMALLINT,
   games_played SMALLINT,
   goals_for SMALLINT,
   goals_against SMALLINT,
   PRIMARY KEY (year, team)
);
```

```
CREATE TABLE fuchibol.Encuentros (
   game_id INTEGER,
   home_team VARCHAR (255) REFERENCES fuchibol.Team(team),
   away_team VARCHAR (255) REFERENCES fuchibol.Team(team),
   date VARCHAR (255),
   year INTEGER,
   time_utc TIME,
   attendance INTEGER,
   venue VARCHAR (255),
   league VARCHAR (255),
   PRIMARY KEY (game_id, home_team, away_team),
   UNIQUE (game_id)
);
```

```
CREATE TABLE fuchibol.Eventos (
      game_id INTEGER REFERENCES fuchibol.Encuentros(game_id),
      minute VARCHAR (255),
      event VARCHAR (255),
   PRIMARY KEY (game_id, minute, event)
);
```

```
COPY team
FROM '/home/cc3201/datos/team.csv'
DELIMITER ',' CSV HEADER;
```

```
COPY encuentros
FROM '/home/cc3201/datos/encuentros.csv'
DELIMITER ',' CSV HEADER;
```

```
COPY stats
FROM '/home/cc3201/datos/stats.csv'
DELIMITER ',' CSV HEADER;
```

```
COPY eventos
FROM '/home/cc3201/datos/eventos.csv'
DELIMITER ',' CSV HEADER;
```

Estado final base de datos:

Team: 42 filas.
Stats: 420 filas.
Encuentros: 7589 filas.
Eventos: 589754 filas.

Todo el procesamiento principal de los datos se hizo con Python (ver anexo).
Primero, se eliminaron todas las columnas excepto las primeras 9 de *matches.csv* dado que es información que no se utilizará. Además, los partidos que no tenían valor en “attendance” se fijaron en 0 y los que no tenían “venue” se eliminaron de la tabla. También se tuvo que eliminar las filas duplicadas y cambiar el formato de los valores de “attendance” puesto que la coma (“,”) generaba error de syntax en Postgre.
Respecto a *events.csv*, se eliminaron las filas que no tuvieran ningún minuto asignado, excepto los que tuvieran un comentario interesante. Además se removieron las que no tenían comentarios útiles como “no commentary” “Event”, “Kickoff” o “End Match”, y las filas de los “game_id” que fueron

eliminados de *matches.csv* puesto que es una llave foránea a la tabla **Encuentros.**
Con *all_tables.csv* el principal problema fue que los nombres de los equipos se encontraban en el formato abreviado de 3 letras, entonces se reemplazaron manualmente en Excel por los del formato en *matches.csv*. Además, a partir de *all_tables.csv* se creó *team.csv* para la tabla **Team.**

```
cc3201=# SELECT * FROM team LIMIT 10;
          team
------------------------
 Arsenal
 Aston Villa
 Brighton & Hove Albion
 Birmingham City
 Blackburn Rovers
 Blackpool
 Bolton Wanderers
 AFC Bournemouth
 Brentford
 Burnley
(10 rows)
```

```
cc3201=# SELECT * FROM stats LIMIT 10;
 year |       team        | place | points | won | lost | drew | games_played | goals_for | goals_against
------+-------------------+-------+--------+-----+------+------+--------------+-----------+---------------
 2001 | Arsenal           |     1 |     87 |  26 |    3 |    9 |           38 |        79 |            36
 2001 | Liverpool         |     2 |     80 |  24 |    6 |    8 |           38 |        67 |            30
 2001 | Manchester United |     3 |     77 |  24 |    9 |    5 |           38 |        87 |            45
 2001 | Newcastle United  |     4 |     71 |  21 |    9 |    8 |           38 |        74 |            52
 2001 | Leeds United      |     5 |     66 |  18 |    8 |   12 |           38 |        53 |            37
 2001 | Chelsea           |     6 |     64 |  17 |    8 |   13 |           38 |        66 |            38
 2001 | West Ham United   |     7 |     53 |  15 |   15 |    8 |           38 |        48 |            57
 2001 | Aston Villa       |     8 |     50 |  12 |   12 |   14 |           38 |        46 |            47
 2001 | Tottenham Hotspur |     9 |     50 |  14 |   16 |    8 |           38 |        49 |            53
 2001 | Blackburn Rovers  |    10 |     46 |  12 |   16 |   10 |           38 |        55 |            51
(10 rows)
```

```
cc3201=# SELECT * FROM eventos LIMIT 10;
 game_id | minute |                    event
---------+--------+---------------------------------------------
   18432 | 4'     | Darius Vassell Goal
   18432 | 45'    | On: Phil Neville|Off: Mikael Silvestre
   18432 | 45'    | Halftime
   18432 | 52'    | Juan Sebastián Verón Yellow Card
   18432 | 64'    | On: Moustapha Hadji|Off: Paul Merson
   18432 | 70'    | On: Andrew Cole|Off: David Beckham
   18432 | 72'    | Paul Scholes Yellow Card
   18432 | 78'    | On: Bosko Balaban|Off: Juan Pablo Angel
   18432 | 83'    | On: Ole Gunnar Solskjaer|Off: Paul Scholes
   18432 | 90'    | Ozalan Alpay (OG)
(10 rows)
```

```
cc3201=# SELECT * FROM eventos LIMIT 10;
 game_id | minute |                  event
---------+--------+-------------------------------------------
   18432 | 4'     | Darius Vassell Goal
   18432 | 45'    | On: Phil Neville|Off: Mikael Silvestre
   18432 | 45'    | Halftime
   18432 | 52'    | Juan Sebastián Verón Yellow Card
   18432 | 64'    | On: Moustapha Hadji|Off: Paul Merson
   18432 | 70'    | On: Andrew Cole|Off: David Beckham
   18432 | 72'    | Paul Scholes Yellow Card
   18432 | 78'    | On: Bosko Balaban|Off: Juan Pablo Angel
   18432 | 83'    | On: Ole Gunnar Solskjaer|Off: Paul Scholes
   18432 | 90'    | Ozalan Alpay (OG)
(10 rows)
```

Consultas:

¿Cuántas victorias, derrotas y empates ha tenido cada equipo en la 2003-2004 Barclays Premier League?

```sql
SELECT team, SUM(won) AS victorias, SUM(lost) AS derrotas, SUM(drew) AS
empates
FROM fuchibol.Stats s
JOIN fuchibol.Encuentros e ON s.team = e.home_team OR s.team =
e.away_team
WHERE league = '2002-2003 Barclays Premier League'
GROUP BY team;
```

¿Cuál es el promedio de goles por partido para cada equipo en la 2003-2004 Barclays Premier League en el año 2010?

```sql
SELECT t.team, AVG(gf) AS promedio_goles
FROM fuchibol.team t
JOIN fuchibol.encuentros e ON t.team = e.home_team OR t.team =
e.away_team
JOIN fuchibol.stats s ON t.team = s.team AND e.year = s.year
WHERE e.league = '2003-2004 Barclays Premier League' AND e.year = 2010
GROUP BY t.team;
```

¿Cuál ha sido el mayor porcentaje de victorias que ha tenido everton y cuál fue el año?

```
SELECT s.year, s.team, (s.won::numeric / s.games_played) * 100 AS
porcentaje_victorias
FROM fuchibol.stats s
WHERE s.team = 'Everton'
ORDER BY porcentaje_victorias DESC
LIMIT 1;
```

Índices:

- **Team:**

```
CREATE INDEX teamindex ON fuchibol.Team (team);
```

- **Stats:**

```
CREATE INDEX statsindex ON fuchibol.Stats (year, team);
```

- **Encuentros:**

```
CREATE INDEX encuentrosindex ON fuchibol.Encuentros (game_id, home_team,
away_team);
```

- **Eventos:**

```
CREATE INDEX eventosindex ON fuchibol.Eventos (game_id, minute, event);
```

Anexo:
[Código para la limpieza de .csv's]