

# 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**.

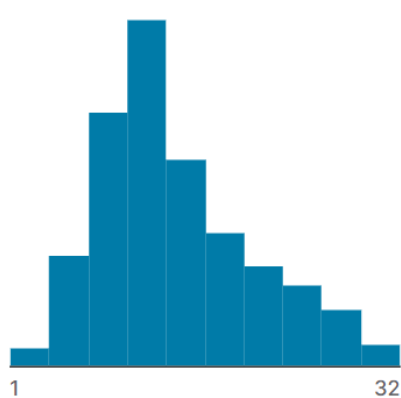
Exploración de Datos:

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

## Distribución de “W” de all\_tables.csv:

# W

Games Won

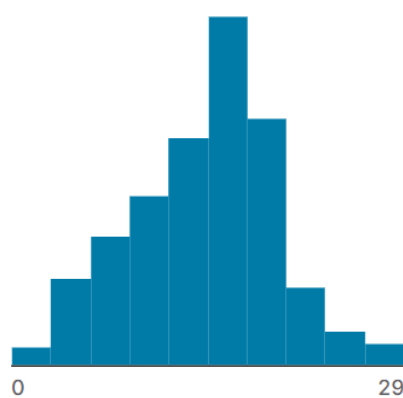


Valid	■	420	100%		
Mismatched	■	0	0%		
Missing	■	0	0%		
Mean			14.3		
Std. Deviation			6.01		
Quantiles			1	Min	
			10	25%	
			13	50%	
			18	75%	
			32	Max	

## Distribución de “L” de all\_tables.csv:

# L

Games Lost

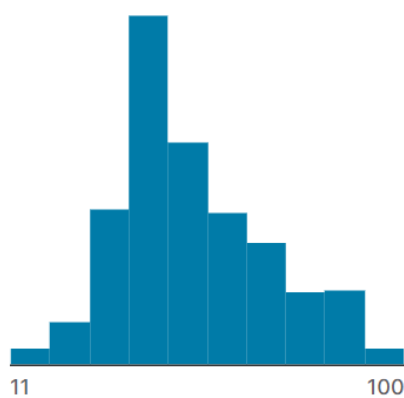


Valid	■	420	100%		
Mismatched	■	0	0%		
Missing	■	0	0%		
Mean			14.3		
Std. Deviation			5.58		
Quantiles			0	Min	
			10	25%	
			15	50%	
			18	75%	
			29	Max	

## Distribución de “Points” de all\_tables.csv:

# P

Points

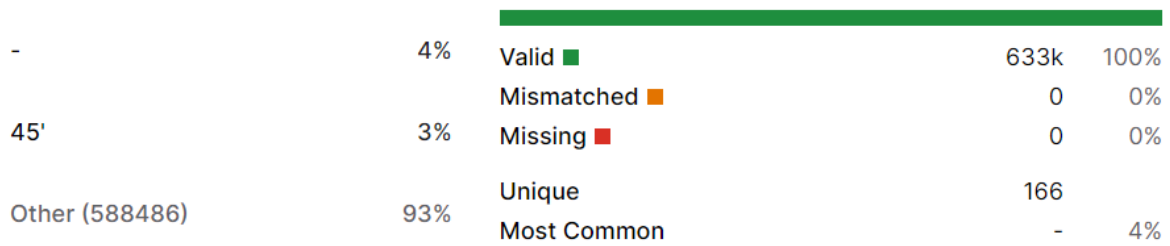


Valid	■	420	100%		
Mismatched	■	0	0%		
Missing	■	0	0%		
Mean			52.2		
Std. Deviation			17.2		
Quantiles			11	Min	
			40	25%	
			48	50%	
			64	75%	
			100	Max	

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

### A Time

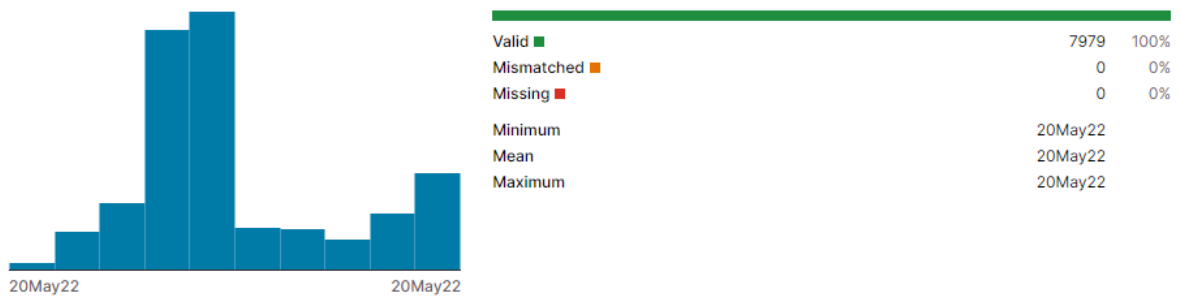
time in match



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

### time (utc)

start of match time



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

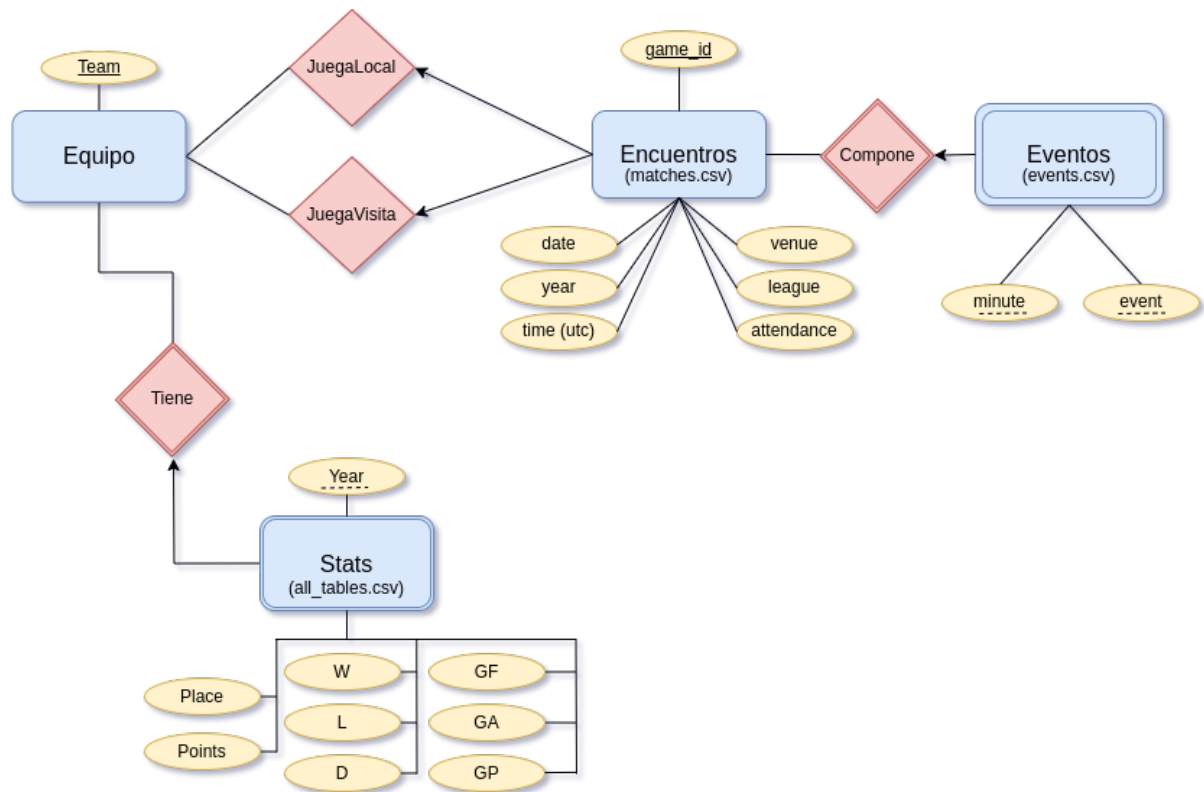
<b>Atributos events.csv</b>	<b>Dominio</b>
id	Int
Time	String
Event	String

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

Consultas (queries) 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 encuentros LIMIT 10;
```

game_id	home_team	away_team	date	year	time_utc	attendance	venue	league
18123	Charlton Athletic	Everton	Saturday, August 18	2001	14:00:00	20451	The Valley, London, England	2001-2002 Barclays Premier League
18124	Derby County	Blackburn Rovers	Saturday, August 18	2001	14:00:00	28236	Pride Park Stadium, England	2001-2002 Barclays Premier League
18125	Leeds United	Southampton	Saturday, August 18	2001	14:00:00	39715	Elland Road, Leeds, England	2001-2002 Barclays Premier League
18126	Leicester City	Bolton Wanderers	Saturday, August 18	2001	14:00:00	19987	Filbert Street, England	2001-2002 Barclays Premier League
18127	Liverpool	West Ham United	Saturday, August 18	2001	14:00:00	43935	Anfield, Liverpool, England	2001-2002 Barclays Premier League
18128	Middlesbrough	Arsenal	Saturday, August 18	2001	14:00:00	31557	The Riverside Stadium, England	2001-2002 Barclays Premier League
18129	Sunderland	Ipswich Town	Saturday, August 18	2001	14:00:00	47370	Stadium of Light, Sunderland, England	2001-2002 Barclays Premier League
18130	Tottenham Hotspur	Aston Villa	Saturday, August 18	2001	14:00:00	36059	White Hart Lane, England	2001-2002 Barclays Premier League
18199	Chelsea	Newcastle United	Sunday, August 19	2001	13:00:00	40124	Stamford Bridge, London, England	2001-2002 Barclays Premier League
18200	Manchester United	Fulham	Sunday, August 19	2001	15:00:00	67534	Old Trafford, Manchester, England	2001-2002 Barclays Premier League

(10 rows)

## Consultas (considerando un input particular):

### Victorias, derrotas y empates de cada equipo entre el 2000 y el 2004:

```
SELECT team, SUM(won) AS victorias,
SUM(lost) AS derrotas, SUM(drew) AS empates

FROM fuchibol.stats s
WHERE year BETWEEN 2000 AND 2004
GROUP BY team
ORDER BY team;
```

### Registro de tarjetas rojas y amarillas en el estadio Old Trafford:

```
SELECT
    (SELECT COUNT(*)
     FROM fuchibol.encuentros en, fuchibol.tarjetas_rojas tr
     WHERE en.game_id = tr.game_id
     AND LOWER(venue) LIKE LOWER('%Old Trafford%')) AS rojas,
    (SELECT COUNT(*)
     FROM fuchibol.encuentros en, fuchibol.tarjetas_amarillas ta
     WHERE en.game_id = ta.game_id
     AND LOWER(venue) LIKE LOWER('%Old Trafford%')) AS amarillas;
```

### Los porcentajes de victorias y cantidad de publico por anho de Arsenal:

```
SELECT s.year, s.team, ROUND((s.won / s.games_played::numeric) * 100, 2)
AS porcentaje_victorias, ah.att_home, aa.att away,
(ah.att_home+aa.att_away) AS att_tot
FROM fuchibol.stats s, fuchibol.att_home ah, fuchibol.att_away aa
WHERE s.team = 'Arsenal'
AND ah.home_team = s.team
AND ah.year = s.year
AND aa.away_team = s.team
AND aa.year = s.year
ORDER BY porcentaje_victorias DESC;
```

## Vistas:

```
CREATE MATERIALIZED VIEW tarjetas_rojas AS
SELECT game_id, event
FROM fuchibol.eventos
WHERE LOWER(event) LIKE '%red card%';
```

```
CREATE MATERIALIZED VIEW tarjetas_amarillas AS
SELECT game_id, event
FROM fuchibol.eventos
WHERE LOWER(event) LIKE '%yellow card%';
```

```
CREATE MATERIALIZED VIEW att_away AS
SELECT away_team, year, SUM(attendance) AS att_away
FROM encuentros
GROUP BY away_team, year;
```

```
CREATE MATERIALIZED VIEW att_home AS
SELECT home_team, year, SUM(attendance) AS att_home
FROM encuentros
GROUP BY home_team, year;
```

## Anexo:

[Código para la limpieza de .csv's](#)