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**.

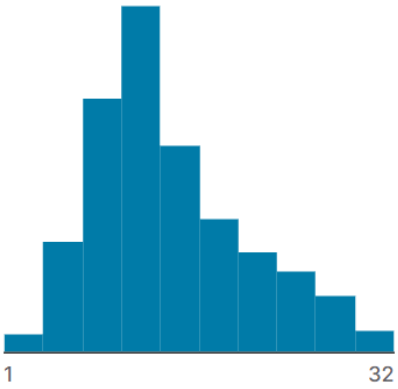
Exploración de Datos:

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

Distribución de “W” de all_tables.csv:

W

Games Won

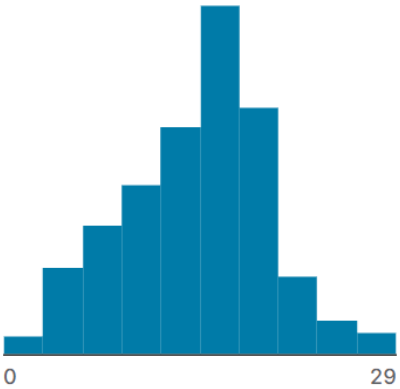


Valid	420	100%
Mismatched	0	0%
Missing	0	0%
Mean	14.3	
Std. Deviation	6.01	
Quantiles	1	Min
	10	25%
	13	50%
	18	75%
	32	Max

Distribución de “L” de all_tables.csv:

L

Games Lost

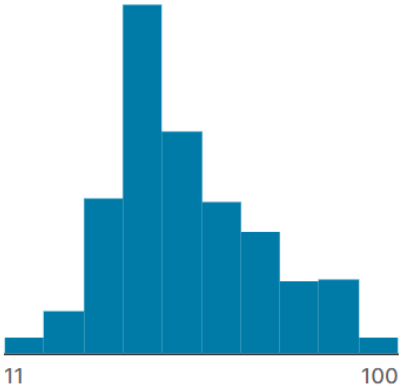


Valid	420	100%
Mismatched	0	0%
Missing	0	0%
Mean	14.3	
Std. Deviation	5.58	
Quantiles	0	Min
	10	25%
	15	50%
	18	75%
	29	Max

Distribución de “Points” de all_tables.csv:

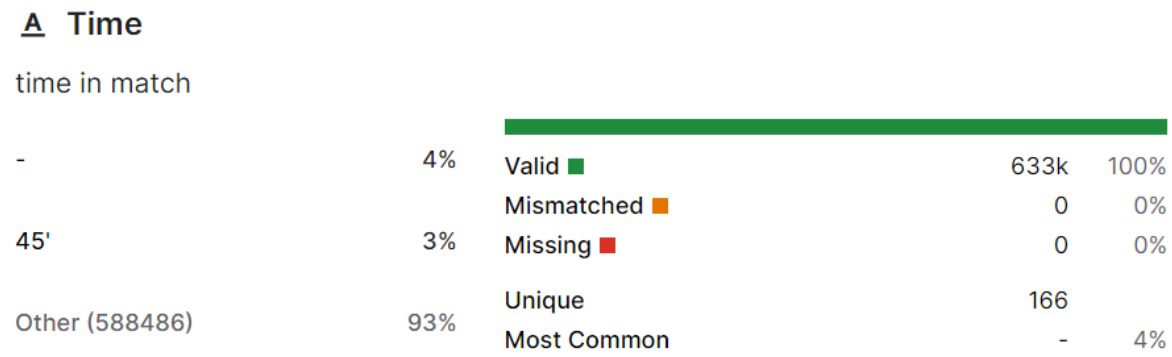
P

Points

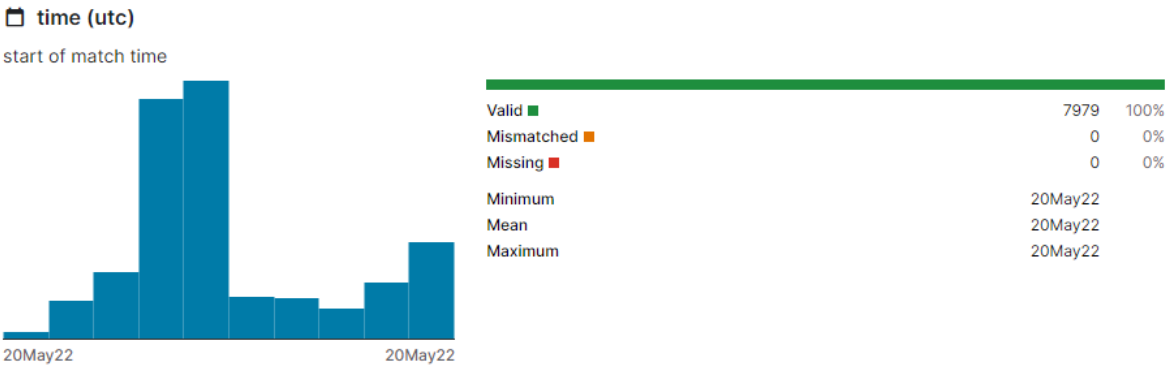


Valid	420	100%
Mismatched	0	0%
Missing	0	0%
Mean	52.2	
Std. Deviation	17.2	
Quantiles	11	Min
	40	25%
	48	50%
	64	75%
	100	Max

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

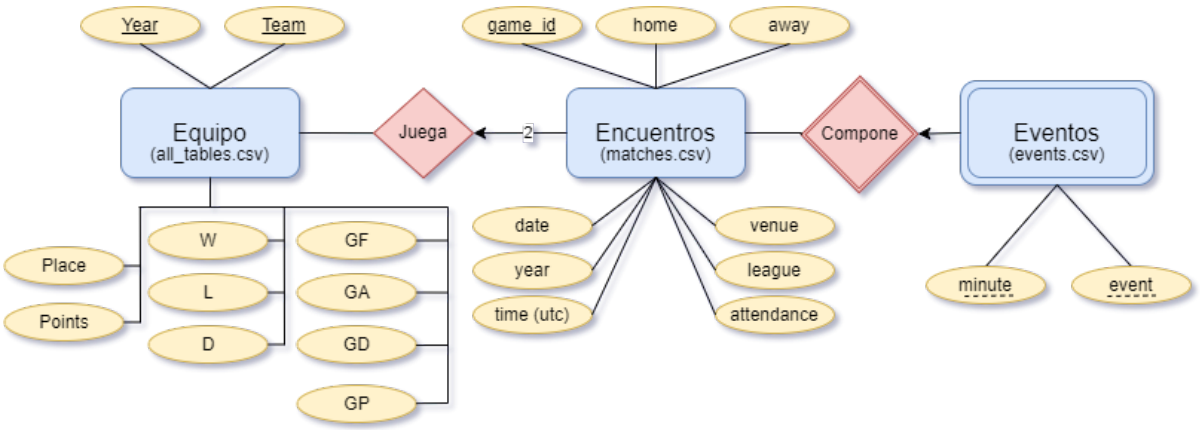
Atributos events.csv	Dominio
id	Int
Time	String
Event	String

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:

Equipo(year, team, place, points, w, l, d, gp, gf, ga, gd)
Encuentros(game_id, home, away, date, year, time_utc, venue, league, attendance)
Eventos(En.game_id, minute, event)
Juega(En.game_id, Eq.year, Eq.team)

¿2NF?¿3NF?¿BCNF?:

Todas las tablas están al menos en 2NF, dado que no hay dependencias funcionales parciales. Sin embargo, la tabla **Equipo** podría considerarse fuera de 3NF puesto que gd (*goal difference*) resulta de la diferencia entre gf y ga (*goal for* y *goal against*). Entonces, como {year, team} → {gf, ga} → {gd}, no se cumpliría 3NF por dependencia transitiva.
Encuentros, **Eventos** y **Juega** están en BCNF. No hay dependencias transitivas y se cumple que la dependencia funcional viene desde la super llave.