DEPARTAMENTO DE COMPUTACION

Facultad de Ciencias Exactas y Naturales - UBA

Trabajo Práctico 1

Consultas SQL, DER, Modelo Relacional, Análisis Exploratorio, Visualizaciones

3 de noviembre de 2024 — Laboratorio de datos

Grupo 1

Integrante	LU	Correo electrónico
Espínola, Marcos Nahuel	827/18	espinola.marcos98@gmail.com
Fernández, Ezequiel Juan	774/23	ezequieljuanfernandez2003@gmail.com
Herrera, Alison Yamila	814/23	yaliherrera02@gmail.com

Facultad de Ciencias Exactas y Naturales
Universidad de Buenos Aires
Ciudad Universitaria - (Pabellón I/Planta Baja)
Intendente Güiraldes 2610 - C1428EGA
Ciudad Autónoma de Buenos Aires - Rep. Argentina
Tel/Fax: (++54 +11) 4576-3300
http://www.exactas.uba.ar

1. Primer Entregrable: Problema a modelar

1.1. Diseño de nuestro DER

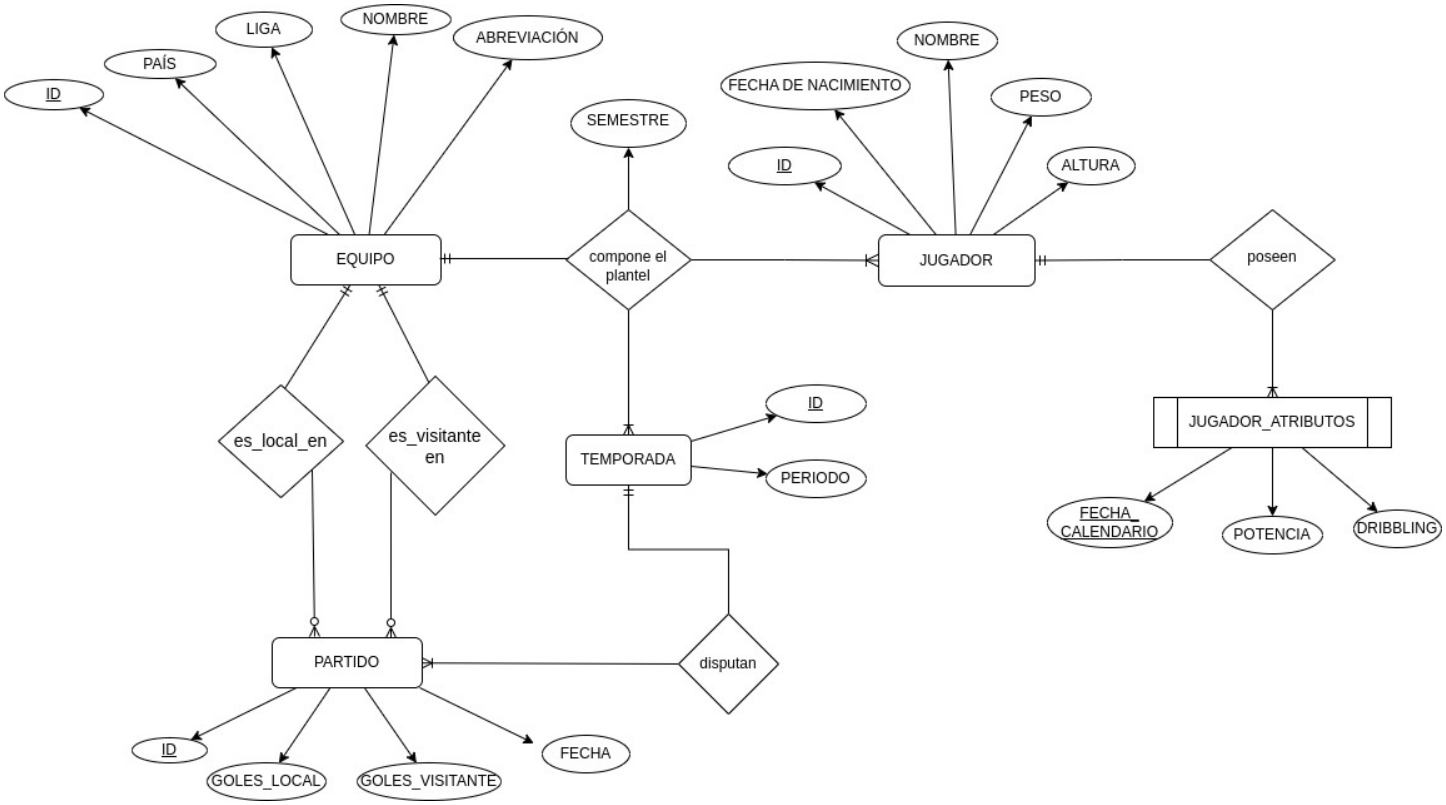


Figura 1

Tomamos, FECHA como la jornada de la liga, tipo fixture. Y FECHA_CALENDARIO con el formato (dd/mm/aa).

1.2. Síntesis de las entidades

Según pudimos interpretar de la consigna del trabajo práctico, las **entidades** que identificamos fueron las siguientes:

1. **Jugador**: Contiene la nómina de todos los jugadores que participan en nuestro modelado.
2. **Jugador_Atributos**: Describe las características físicas del jugador según la fecha calendario en la que fueron tomados estos datos. Dejamos los atributos fecha, dribbling y potencia ya que es información que nos va a servir para la parte del análisis.
3. **Equipo**: Contiene la nómina de todos los equipos involucrados en las ligas y en el período de tiempo definido en Temporada.
4. **Temporada**: Dada la relevancia que tiene en nuestro modelado la dimensión temporal, ya sea para definir de manera unívoca los planteles de los equipos, así como también los partidos que se disputan, se incluye en nuestro trabajo esta entidad para poder complementar la información de las demás entidades. De esta manera, podemos ubicar en el tiempo la información en nuestro modelo entre las temporadas 2008/2009 y la 2015/2016.
5. **Partido:** Contiene el historial de todos los partidos jugados de todas las ligas entre los rangos de temporadas definidos por la entidad Temporada.

Explicación de las relaciones del DER:

1. **Componen_plantel**: Es una relación ternaria, que relaciona jugador, temporada y equipo, para saber en qué club está jugando cada jugador y cuándo lo está haciendo.
2. **Poseen**: Es una relación identificativa, relaciona Jugador y Jugador_Atributos para saber sobre qué atributos de qué jugador estamos hablando.
3. **Es_visitante_en**: Relaciona un equipo con un partido a partir de la localía que tenga en ese encuentro, que en este caso deberá ser visitante.
4. **Es_local_en**: Relaciona un equipo con un partido a partir de la localía que tenga en ese encuentro, que en este caso deberá ser local.
5. **Disputan:** Relaciona partido y temporada, a partir de esta relación es que puede saber cuando los partidos son jugados y así identificarlos (teniendo en cuenta el atributo fecha de partido).

2. Segundo Entregable: Modelo Relacional y Normalización

ENTIDADES FUERTES

- JUGADOR (ID, FECHA_DE_NACIMIENTO, NOMBRE, PESO, ALTURA)
- EQUIPO (ID, NOMBRE, ABREVIACIÓN, PAÍS, LIGA)
- PARTIDO (ID, ID _TEMPORADA, ID_EQUIPO _LOCAL, ID_EQUIPO_VISITANTE, GOLES_LOCAL, GOLES _VISITANTE, FECHA)

  Por tener una cardinalidad de 1:N toma como foreign key las claves primarias de la relación con TEMPORADA y EQUIPO

- TEMPORADA (ID, PERIODO)

ENTIDADES DEBILES

- JUGADOR_ATRIBUTOS (ID_JUGADOR, FECHA_CALENDARIO, POTENCIA, DRIBBLING)

  (ID_JUGADOR ES FK Y forma parte de la PK POR SER ENTIDAD DÉBIL).
  Explicación igual que la anterior, por tener cardinalidad 1:N

RELACION 1:N

- DISPUTAN (ID_TEMPORADA, ID_PARTIDO)
- ES _LOCAL _EN (ID_EQUIPO, ID_PARTIDO)
- ES _VISITANTE _EN (ID_EQUIPO, ID_PARTIDO)

RELACION TERNARIA

- COMPONEN _EL_PLANTEL (ID_JUGADOR, ID_TEMPORADA, ID_EQUIPO)

LAS PALABRAS QUE ESTÁN SUBRAYADAS SON PK
LAS PALABRAS QUE ESTAN EN ROJO SON FK

EQUIPO
Contamos con las siguientes dependencias funcionales:

Df1: { Id_equipo } → {Nombre, País}

Df2: {Nombre} → {Abreviación}

Df3: {País} → {Liga}

Se encuentra en primera forma normal porque los valores de los atributos son atómicos. Se encuentra en segunda forma normal porque como tenemos una única clave primaria (Id) no van a existir dependencias parciales. No se encuentra en tercera forma normal porque existen dependencias transitivas entre ID → ABREVIACIÓN y ID → *LIGA*.
Para llevarlo a la tercera forma normal, debemos descomponer en 3 tablas, de acuerdo a las dependencias funcionales encontradas. Las llamamos Equipo_A, Equipo _B, Equipo_C

JUGADOR
Df1: {Id} → {Nombre, Fecha de nacimiento, Altura}
Se encuentra tanto en primera como segunda y tercera forma normal

JUGADOR_ATRIBUTOS
Df1: {Id_jugador, Fecha_calendario } → {Potencia, Dribbling}
Tanto potencia como dribbling dependen de manera completa de la pk (Id_jugador, Fecha).
Se encuentra tanto en primera, como segunda y tercera forma normal.

TEMPORADA
Df1: {Id} → {Periodo}
Se encuentra tanto en primera como segunda y tercera forma normal

PARTIDO
Contamos con las siguientes dependencias funcionales
Df1: { Id_partido } $\rightarrow \{Id\_temporada, Fecha, Id\_equipo\_local, Id\_equipo\_visitante\}$
Df2: {Id_temporada, Fecha,Id_equipo_local} $\rightarrow$
$\{Id\_equipo\_visitante, Goles\_local, Goles\_visitante Id\_equipo\_local, Id\_equipo\_visitante\}$

Se encuentra en primera forma normal porque los valores de los atributos son atómicos.
Se encuentra en segunda forma normal porque como tenemos una única clave primaria (Id) no van a existir dependencias parciales.
No se encuentra en tercera forma normal porque existen dependencias transitivas entre Id_partido $\rightarrow$ $Id\_equipo\_Visitante, Goles\_local, Goles\_visitante$ por ejemplo.
Lo que hicimos fue considerar las dependencias funcionales 1 y 2 para hacer dos tablas PARTIDO_A y PARTIDO_B, para eliminar la transitividad. LA DF 2 es la dependencia que toma menos atributos en su conjunto de llegada para poder determinar a los de salida.
Para la etapa de análisis consideraremos a PARTIDO en segunda forma normal

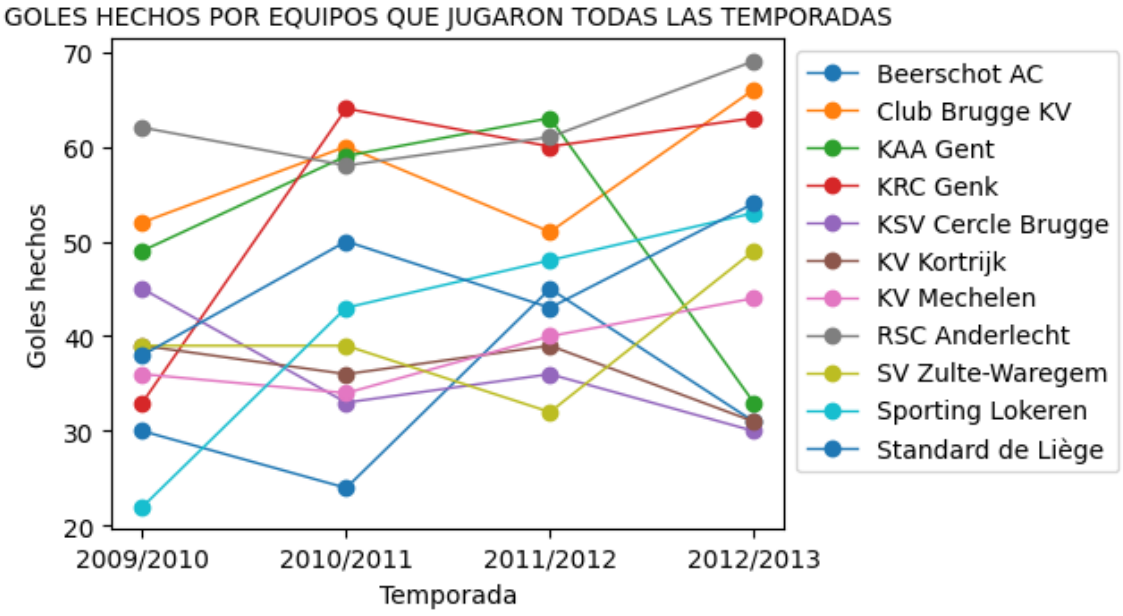
3. Análisis Exploratorio: Consultas y visualizaciones

Para el análisis, necesitábamos seleccionar cuatro años consecutivos de datos. Al revisar las temporadas disponibles (2007-2016), encontramos que en 2007/2008 apenas se registraron atributos de los jugadores, presentando una gran cantidad de valores nulos. En la temporada 2008/2009, hubo una notable falta de datos sobre los jugadores que participaron en los partidos, lo cual limitaría las consultas SQL y los análisis futuros. Por otro lado, en la temporada 2013/2014, la liga belga tuvo una pausa o no registró todos los partidos, ya que solo se documentaron 6 partidos.

Por estas razones, decidimos utilizar el período de 2009 a 2013, abarcando así cuatro temporadas consecutivas (2009/2010 - 2012/2013) en las que los datos necesarios están presentes en su mayoría. Aunque existen algunos valores nulos en estas temporadas, son demasiado pocos para afectar significativamente el análisis.

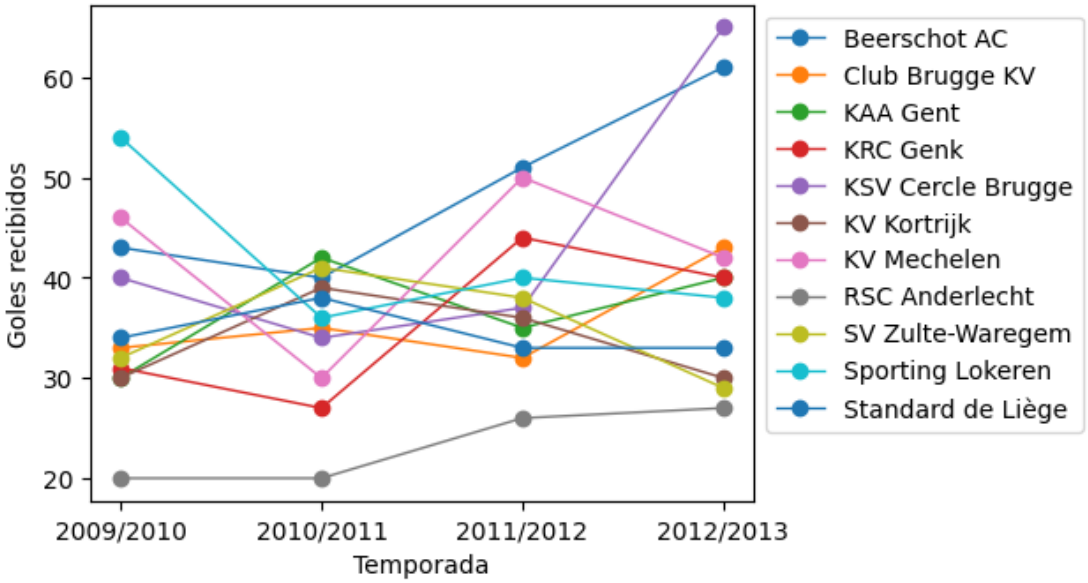
3.1. Graficar la cantidad de goles a favor y en contra de cada equipo a lo largo de los años que elijan.

Nosotros decidimos separar el gráfico entre goles recibidos y goles hechos, a su vez los dividimos entre los equipos que habían jugado todas todas las temporadas y los que no.



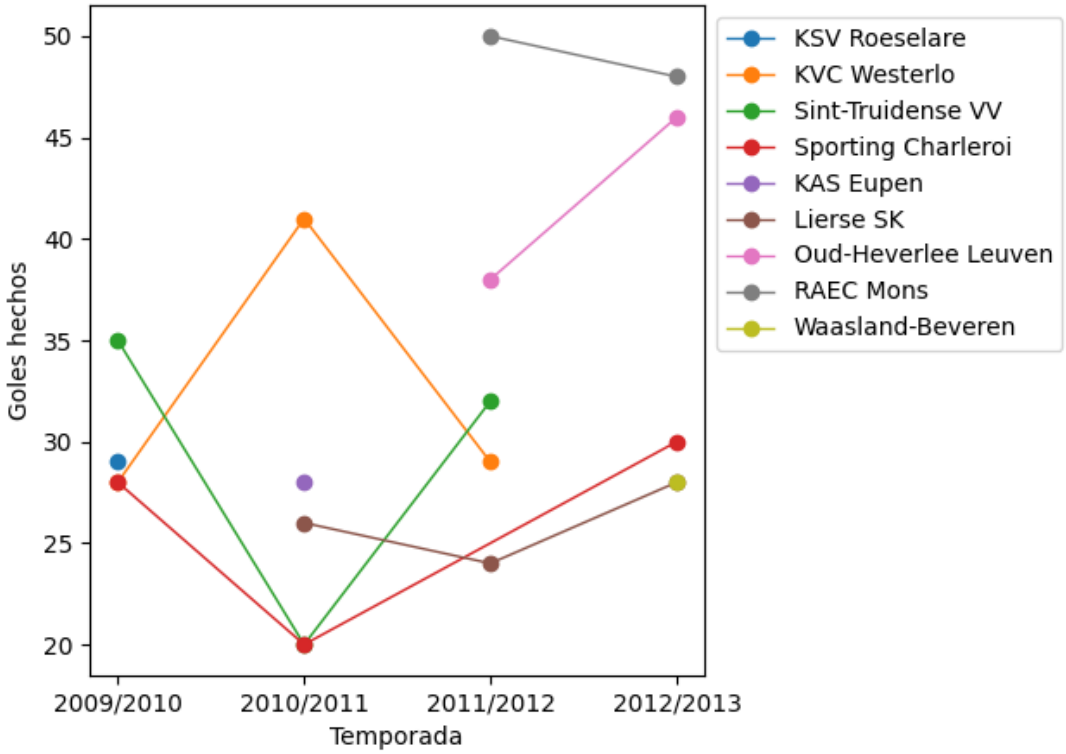
Standar de Liege arranca y termina mas arriba.
Beerschot AC arranca y termina mas abajo.

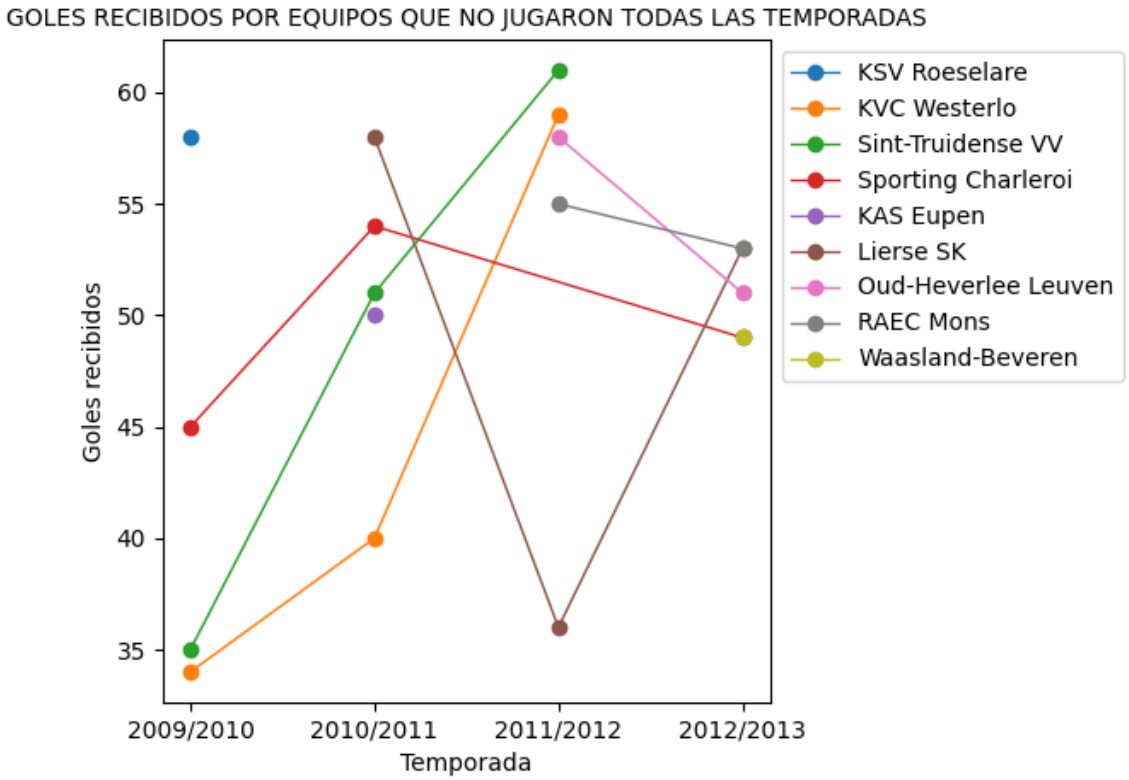
GOLES RECIBIDOS POR EQUIPOS QUE JUGARON TODAS LAS TEMPORADAS



Standar de Liege arranca y termina mas abajo.
Beerschot AC arranca y termina mas arriba.

GOLES HECHOS POR EQUIPOS QUE NO JUGARON TODAS LAS TEMPORADAS





Los equipos que jugaron todas las temporadas eran los que más goles habían metido en total. La mayoría de estos equipos hacían entre 30 y 60 goles y recibían entre 28 y 45.

Los equipos que no jugaron todas las temporadas metieron menos goles; la mayoría de estos equipos metieron entre 20 y 35 goles y recibieron entre 45 y 60.

Podemos observar algunos patrones en el gráfico:

- Los equipos que tienen picos muy altos de goles recibidos junto con picos muy bajos de goles hechos suelen descender.
- En el gráfico, todos los equipos que descienden son aquellos que menos goles hacen y a quienes les hacen más goles en una misma temporada.
- En la última temporada, equipos como el KSV Cercle Brugge y el Beerschot AC, que generalmente son muy goleadores, experimentaron una sequía de goles acompañada de un gran número de goles en contra (son los dos equipos que más goles recibieron esa temporada, con amplia diferencia). Esto podría llevarlos a un descenso al finalizar la temporada.
- El RSC Anderlecht fue el equipo que menos goles recibió en todas las temporadas y el más goleador en dos de ellas. Además, en otra temporada fue el segundo equipo más goleador, y en otra quedó en el cuarto puesto. Esto indica que el RSC Anderlecht es un equipo muy completo que compitió en todas las ligas con posibilidades de ganar el título.

Los descensos están indicados en el gráfico con puntos descontinuados.

3.2. Graficar el promedio de gol de los equipos a lo largo de los años que elijan.

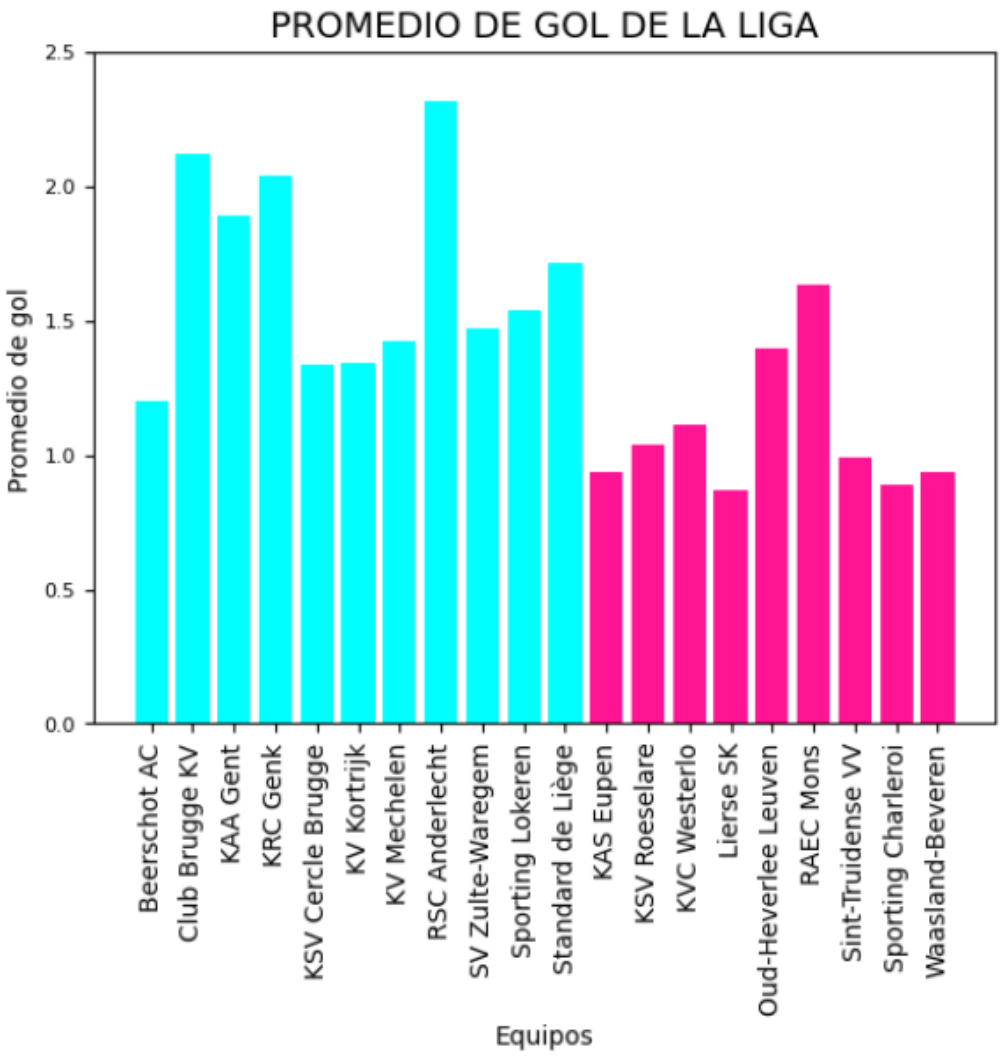


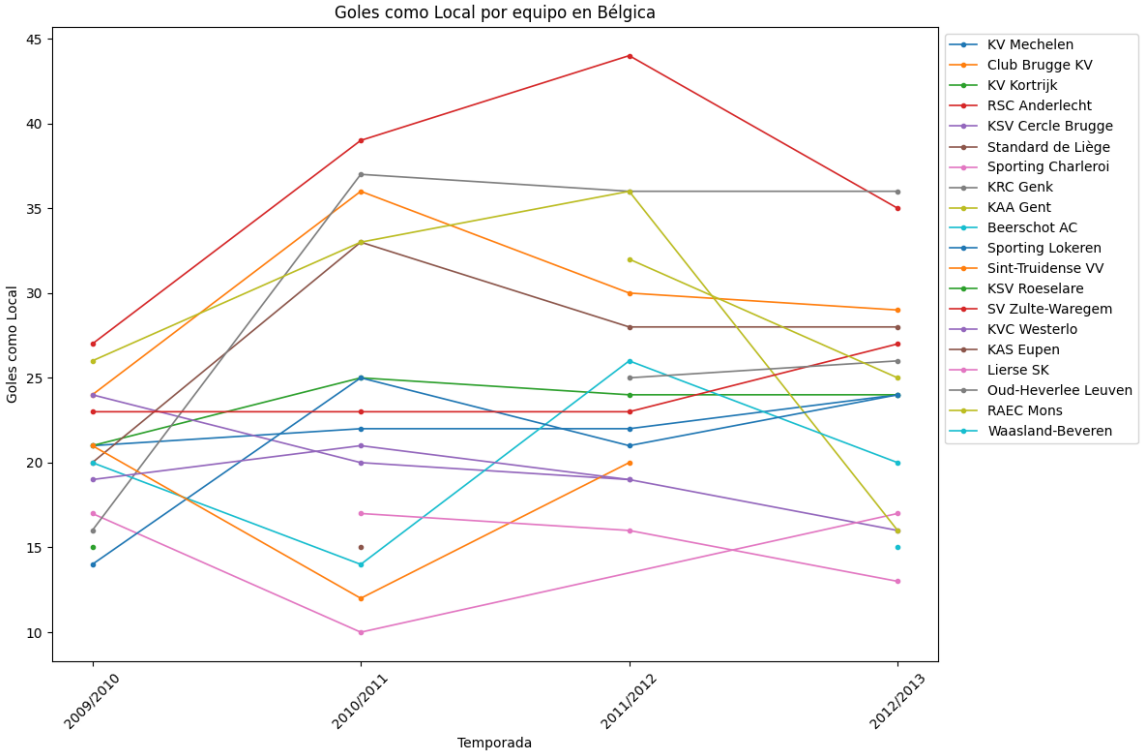
Figura 2: Promedio de gol de la liga a lo largo de los años 2009-2013

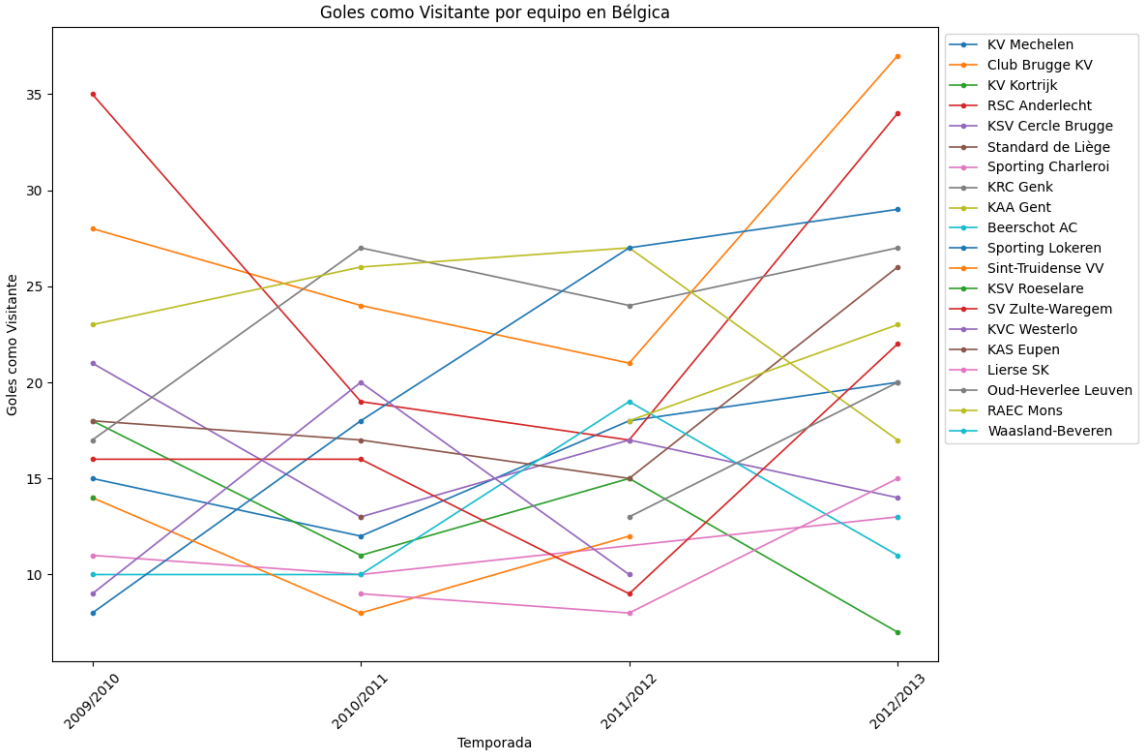
Elegimos este gráfico ya que nos pareció que era muy claro a la hora de visualizar y comparar el promedio de gol de cada uno de los equipos de la liga belga, está separado por los equipos que participaron todas las temporadas y los que no lo hicieron.

Los equipos que participaron de todas las temporadas en el periodo de tiempo elegido son equipos que son más estables y que tienen un nivel alto para sostenerse en primera división, esto separado ve ya que los equipos que se mantuvieron en primera todo el periodo de tiempo reflejan promedios más altos en su gran mayoría respecto a los que participaron en parte de este periodo.

El RSC Anderlecht nuevamente está en lo más alto, esta vez en promedio de goles, lo cual es explicado por su regular y buen rendimiento en este periodo.

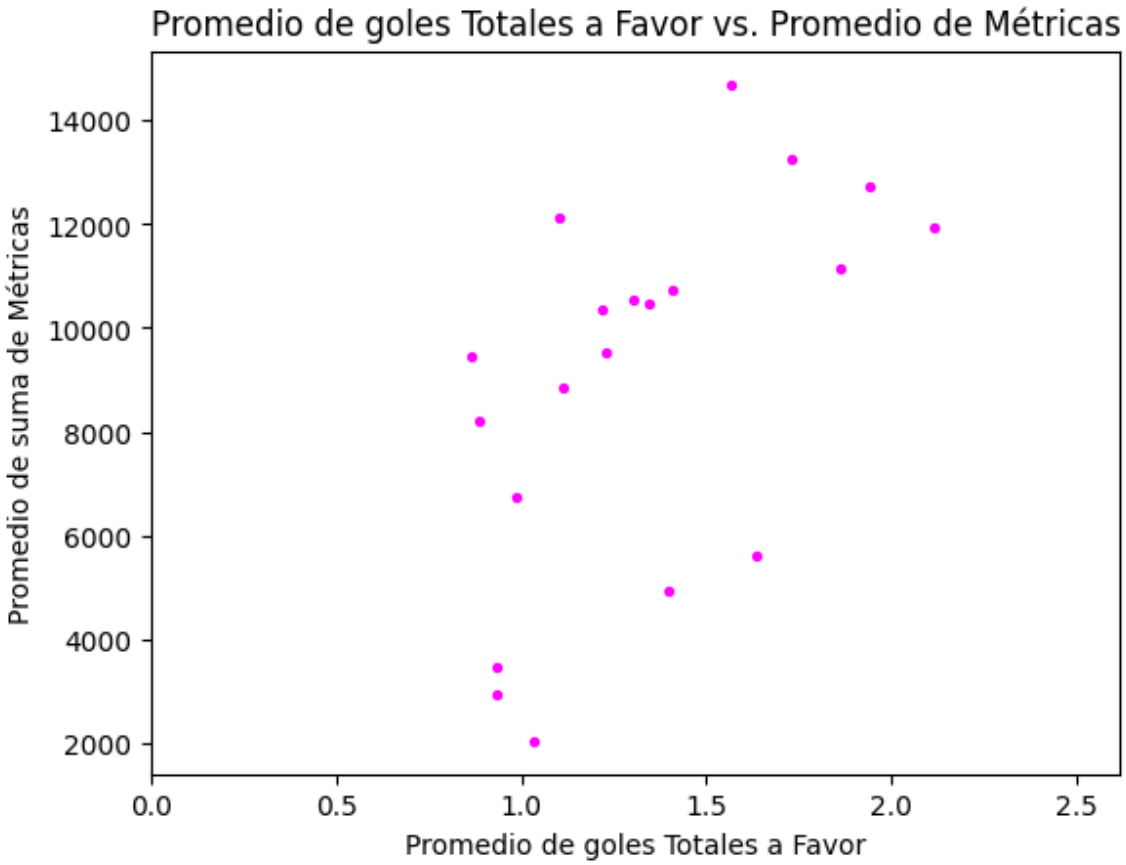
3.3. Graficar la diferencia de goles convertidos jugando de local vs visitante a lo largo del tiempo





Podemos ver que a los equipos les resulta mucho más fácil convertir de local que de visitante, de visitante muchos equipos no llegan a los 20 goles y que de local la mayoría los supera, incluso al equipo equipo con mejor rendimiento en las ligas (por el análisis anterior) el RSC Anderlecht se aprovecha mucho de esto y en su casa se hace muy fuerte, y de visitante tuvo algunas temporadas muy buenas y otras más convencionales (en estas temporadas compensó haciendo muchísimos goles de local).

3.4. Graficar el número de goles convertidos por cada equipo en función de la suma de todos sus atributos.



Decidimos hacer un gráfico de tipo scatterplot para poder observar cómo es que las métricas de un equipo pueden inferir en la cantidad de goles hechos por los equipos. Por lo que observamos que hay una clara tendencia a que, los jugadores con mayores métricas tienen mayor cantidad de goles realizados en los equipos, es decir que existe una relación directamente proporcional donde un equipo con jugadores de mayor calidad tiende a hacer más goles.

Para tener una medida uniforme, tomamos el promedio de la suma de los atributos de los jugadores, específicamente en potencia y dribbling. Esto significa que, si en un año se realizaron varias mediciones, calculamos el promedio de estas para obtener una única medida representativa de cada jugador en la temporada. Luego, sumamos estas métricas a nivel de equipo. A su vez, en el gráfico tomamos el promedio de goles hechos por cada equipo según la cantidad de partidos jugados a lo largo de las temporadas 2009-2013, normalizando así los valores para equiparar de manera equitativa a los equipos que no hayan participado en todas las temporadas. Esto nos permite una comparación justa entre equipos, sin que se vean favorecidos aquellos que jugaron más partidos.