Parcial 1

Tomas Silva
Juan Pabon Vargas

Pontificia Universidad
JAVERIANA
Colombia

Pontificia Universidad Javeriana

3 de Septiembre 2025

Índice:

Presentación

El presente documento data de un informe de los procedimientos, pruebas y análisis realizados durante los acontecimientos del primer parcial de la materia “Procesamiento de datos a gran escala” para el semestre 2530 año 2025. Este informe detalla los procesos de análisis de datos realizados mediante un “serverless compute unit”. Se plantea inicialmente un problema con contexto basado en jugadores de fútbol y sus resultados para su oportuna elección, esta problemática se explicará más a fondo durante el ***planteamiento del problema***, tras esto, se discute la estrategia implementada de análisis para la solución, junto con revisiones de datos, variables, y la calidad de los resultados y análisis junto a elementos de interés.

Introducción

Con el propósito de contextualizar, y como contexto otorgado vía el enunciado del parcial, mi compañero y yo hemos sido contratados por una compañía de fútbol para la creación de un nuevo equipo de fútbol en la liga inglesa. Se tiene el interés de conocer un poco más los equipos y resultados que se obtuvieron en la liga del 17-18, con tal de tener una mejor idea de que tipos de jugadores son preferibles a contratar y que estilo de juego se quiere para su equipo. Para esto contamos con la información de resultados por partido, como también las estadísticas obtenidas por los equipos y jugadores.

Estrategia de Solución

La solución planteada involucra el uso de modelos de predicción para procesamiento con alto volumen de datos, con el fin de analizar las estadísticas de

cada equipo en en liga 17-18 para poder tener mejor conocimiento sobre los equipos y jugadores con tal de poder crear un nuevo equipo de fútbol con los rangos más eficientes según su análisis.

Para analizar los datos, el parcial otorga 3 archivos cargados a la plataforma vía ingestión de datos por volumen:

- Resultados por partidos: conjunto de datos “/csv/resultados_futbol.csv”.
- Estadísticas por equipo: conjunto de datos “/json/temporadas.json”.
- Estadística de jugadores: conjunto de datos “/csv/jugadores.csv”

Estos archivos poseen información respecto a los jugadores correspondientes, sus equipos, y partidos de temporada, información que permitirá analizar el estado y estadísticas de cada jugador y sus mejores atributos.

Conjuntos de datos

Considerando los conjuntos de datos mencionados en la ***Estrategia de solución***, a continuación se presentaran cada conjunto a más detalle junto a sus atributos más importantes:

- resultados_futbol.csv: Muestra resultados en un archivo .csv respecto a resultados por partidos, incluyendo elementos como:
    1. Season (Temporada)
    2. DateTime (Fecha y hora)
    3. HomeTeam (Equipo local)
    4. AwayTeam (Equipo visitante)
    5. FTHG (Objetivos en casa a tiempo completo)
    6. FTAG (Goles como visitante en tiempo completo)
    7. FTR (Resultado de tiempo completo)
    8. HTHG (Goles locales en el descanso)
    9. HTAG (Goles locales en el descanso)
    10. HTR (Goles locales en el descanso)
    11. Referee
    12. HS (Fotos del equipo local)
    13. AS (Fotos del equipo visitante)
    14. HST (El equipo local remata a puerta)
    15. AST (Remates del equipo visitante a portería)
    16. HC (Córners del equipo local)
    17. AC (Córners del equipo visitante)
    18. HF (Faltas del equipo local)
    19. AF (Faltas del equipo visitante)
    20. HY (Tarjetas amarillas del equipo local)

21. AY (Tarjetas amarillas del equipo visitante)
22. HR (Tarjetas rojas del equipo local)
23. AR (Tarjetas rojas del equipo visitante)

De estos datos, los considerados **más** pertinentes para el análisis son, según bajo nuestra opinión: FTR, Faltas de ambos equipos, Tarjetas amarillas de ambos equipos y tarjetas rojas de ambos equipos. Se eligieron estos datos, pues son los que definen cuál de los dos equipos de análisis está en mejor estado, considerando su puntaje, faltas, y tarjetas rojas y amarillas, implicando su condición no solo de eficiencia en el campo, sino que además su viabilidad y comportamiento.

- temporadas.json: Muestra resultados en un archivo .json respecto a resultados por equipos en temporadas. Es de suma importancia reconocer que para el correcto funcionamiento de este análisis, este archivo .json se ha convertido en 2 archivos de formato .csv, en los que se encuentran detallados elementos tanto de las temporadas como de los jugadores en las mismas. Cabe recalcar que, considerando la extensa cantidad de atributos incluidos en este archivo, se mostrarán solamente los valores considerados más importantes para el análisis de los datos. Se incluyen 4 valores, siendo estos:
    1. team_name
    2. team_rating
    3. player_name
    4. player_rating

Estos datos son útiles dado que representan la condición en ratings, medida en desempeño y actividad, no solo del jugador individual, sino que además del equipo en el cual se encuentra.

- jugadores.csv: Muestra resultados en un archivo .csv respecto a resultados por jugadores, incluyendo elementos como:
    1. name
    2. club
    3. age
    4. position (posición - ej: mediocentro, línea de defensa, portero etc)
    5. position_cat (categoría de posición - ej: defensas, porteros etc)
    6. market_value (valor de oferta económica para un jugador)
    7. page_views
    8. fpl_value (puntos totales / precio)
    9. fpl_sel (valores basados en “Fantasy Premier League”)
    10. fpl_points (puntos basados en jugadores reales)
    11. region
    12. nationality
    13. new_foreign
    14. age_cat

15. club_id
16. big_club
17. new_signing

De estos datos, los considerados **más** pertinentes para el análisis son, según nuestra opinión: club, market_value, fpl_value, fpl_points y age. Se eligieron estos datos dado a que muestran valores no solo de rendimiento (como es el caso de los valores basados en fpl) sino que además muestra su rol en el equipo, junto a su precio estimado segun su rendimiento y una forma apropiada de combinar tablas

Reporte de variables y calidad de datos

A la hora de ver los conjuntos de datos y su respectiva calidad se trabajarán dos conceptos principalmente, la calidad general de los datos por cada conjunto, y las variables en total consideradas más importantes.

Primero se va evaluar la calidad de datos del conjunto de datos del archivo: resultados_futbol.csv

1. Completitud:

Season	DateTime	HomeTeam	AwayTeam	FTHG	FTAG	FTR	HTHG	HTAG	HTR	Referee	HS	AS	HST	AST	HC	AC	HF	AF	HY	AY	HR	AR
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Tabla 1, Completitud del conjunto Resultados

   a. Como se puede ver en la ***Tabla 1***, cuando se buscó en el conjunto de datos para ver si existían valores nulos, o imposibles, mostró que no había valores nulos en el conjunto de de datos
2. Consistencia:

   a. Analizando y viendo los tipos de datos y ejemplos de los mismos, se concluye que los datos son consistentes. En cada columna, los datos mantienen el mismo formato. Esto facilita el manejo de los datos más adelante.
3. Validación:
   a. Filtrando datos, se logra confirmar que las fechas están en el rango deseado (de la season 2017 a 2018), nombres de equipos que estuvieron en la liga están bien y los datos que le siguen no contienen valores negativos ni nulos. Esto indica que los valores son válidos y están en buen estado.
4. Exactitud:

a. Inspeccionando los resultados de las partidas con las estadísticas publicadas en ESPN sobre los resultados de los partidos que se encuentran en las tablas sql, se puede confirmar la exactitud de los datos pertinentes.

Ahora se va evaluar la calidad de datos del conjunto de datos del archivo: jugadores.csv

5. Completitud:

name	club	age	position	position_cat	market_value	page_views	fpl_value	fpl_sel	fpl_points	region	nationality	new_foreign	age_cat	club_id	big_club	new_signing
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Tabla 2, Completitud del conjunto jugadores

a. Como se puede ver en la ***Tabla 2***, cuando se buscó en el conjunto de datos para ver si existían valores nulos, o imposibles, mostró que no había valores nulos en el conjunto de datos.

6. Consistencia:
   a. Analizando los datos, se concluye que los datos son consistentes. En cada columna, los datos mantienen el mismo formato. Esto facilita el manejo de los datos más adelante.
7. Validación:
   a. Gracias a lo mostrado en la ***Tabla 2***, es posible evidenciar que los datos están completos, sin elementos nulos o negativos, y que además, cuentan con la información suficiente para analizarse y relacionarse con otras tablas en el análisis pertinente.
8. Exactitud:
   a. Inspeccionando a través de estadísticas en ESPN, es posible evidenciar exactitud en los datos de jugadores, priorizando los valores necesarios para el análisis a realizar.

Ahora se va evaluar la calidad de datos del conjunto de datos del archivo: temporadas.json con sus dos archivos csv creados

9. Completitud:

-RECORD 0	
match_id	0
team_id	0
team_name	0
team	0
team_rating	0
date	0
won_corners	16
att_sv_low_centre	269
won_contest	0
total_tackle	0
aerial_lost	0
possession_percentage	0
accurate_pass	0
total_pass	0
total_throws	0
shot_off_target	11
total_offside	134
blocked_scoring_att	69
ontarget_scoring_att	27

-RECORD 0	
match	0
team	0
player_id	0
player_name	0
player_position_value	0
player_position_info	0
player_rating	0
good_high_claim	13330
touches	3276
saves	12974
total_pass	3326
formation_place	0
accurate_pass	3399
aerial_won	7836
aerial_lost	7494
fouls	8651
total_scoring_att	8718
total_tackle	7684
won_contest	9619

Tabla 3, Completitud y valores del conjunto temporadas y valores del conjunto jugadores en temporadas

a. Como se puede ver en la ***Tabla 3***, cuando se buscó en los conjuntos de datos para ver si existían valores nulos, o imposibles, si se

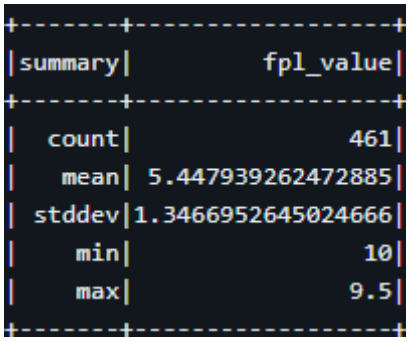
encontraron valores de este tipo en la tabla, por lo que estos datos deberán filtrarse y eliminarse.

10. Consistencia:
    a. Analizando la consistencia de los elementos, se concluye que los datos son consistentes. En cada columna de ambas tablas, los datos mantienen el mismo formato sin contar los nulos existentes. Esto facilita el manejo de los datos más adelante.
11. Validación:
    a. Gracias a lo mostrado en la ***Tabla 3***, es posible evidenciar que los datos están poseen nulos o negativos en algunos de sus casos, valores que deben filtrarse para evitar problemáticas, fuera de esto, los valores son consistentes para el análisis que se mostrará en la sección que prosigue.

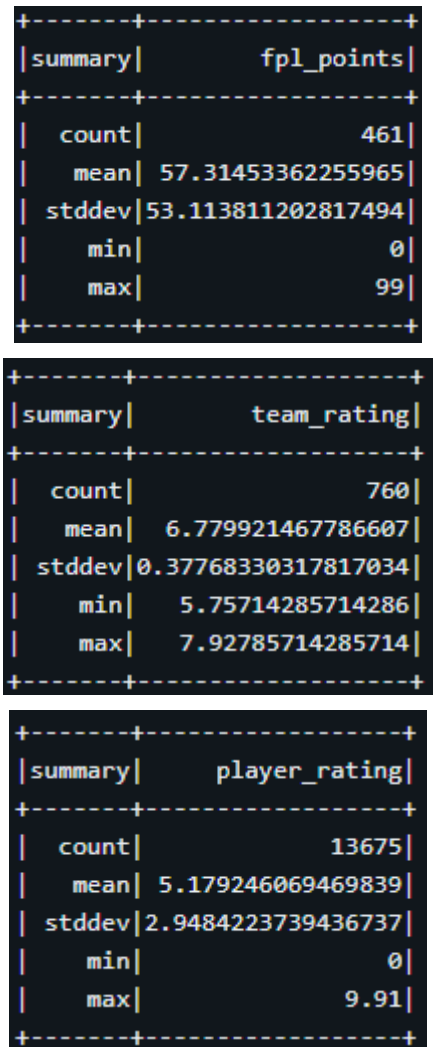
Resultados y análisis

Con tal de encontrar resultados pertinentes, que por igual sean claros y logren mostrar resultados efectivos y útiles, se optó por realizar gráficos específicos para su posterior análisis.

Con tal de obtener los mejores resultados, y con eso, las opciones más viables para la elección de jugadores para el nuevo equipo, se inició recopilando los datos en una sola tabla de gran tamaño, que luego sería filtrada y reducida según los mejores valores. Para esto, inicialmente se obtuvieron los valores promedio de algunos de los valores que se requieren para el análisis, como se ve en la ***Imagen 1*** , conteniendo algunos ejemplos.



summary	fpl_value
count	461
mean	5.447939262472885
stddev	1.3466952645024666
min	10
max	9.5



summary	fpl_points
count	461
mean	57.31453362255965
stddev	53.113811202817494
min	0
max	99

summary	team_rating
count	760
mean	6.779921467786607
stddev	0.37768330317817034
min	5.75714285714286
max	7.92785714285714

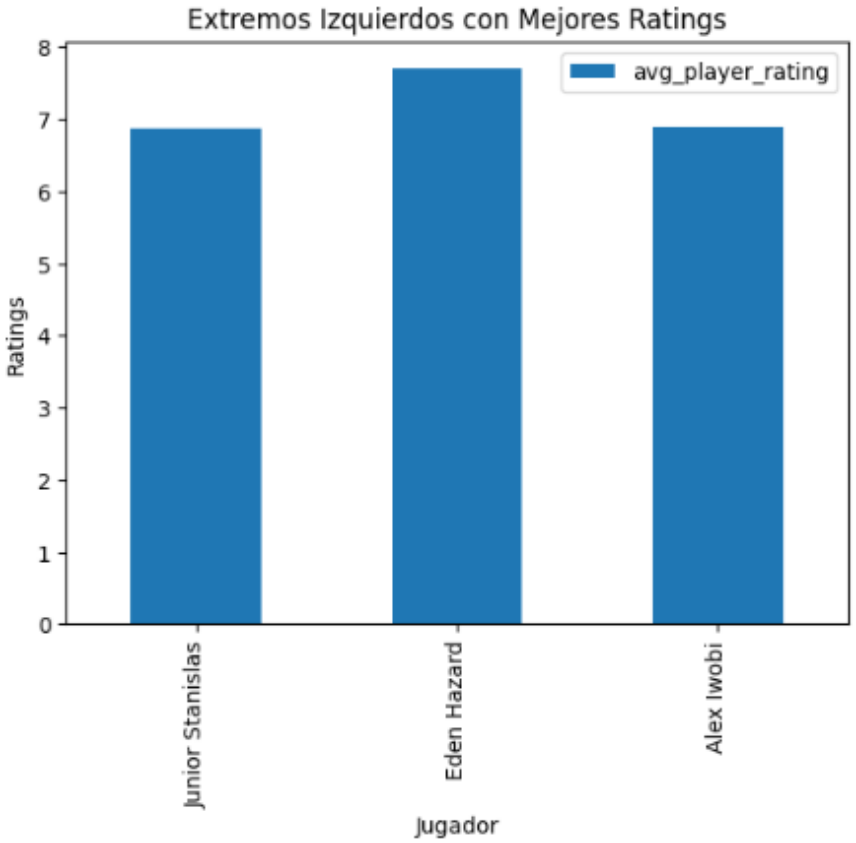
summary	player_rating
count	13675
mean	5.179246069469839
stddev	2.9484223739436737
min	0
max	9.91

Imagen 1, promedios ejemplo de tablas reunidas

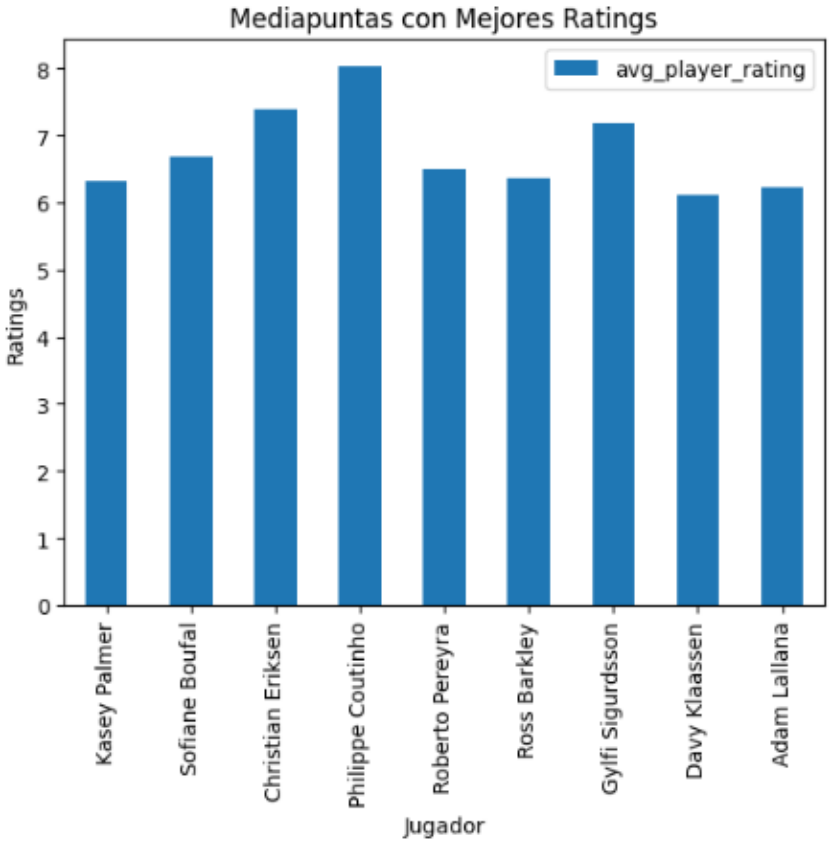
Seguido de ello, se convirtieron los datos para analizarlos, tras lo que se crearon vistas que contengan solamente valores por encima de estos promedios, asegurando la mayor calidad posible (cabe aclarar que en el caso de la edad de los jugadores, es preferible contar con valores menores al promedio de edad, y por lo tanto este requisito también fue filtrado).

Finalmente, y con los valores reunidos de forma pertinente (unión vía fecha, nombre y equipo), se crean gráficas para mostrar los resultados. Estas gráficas se muestran a continuación.

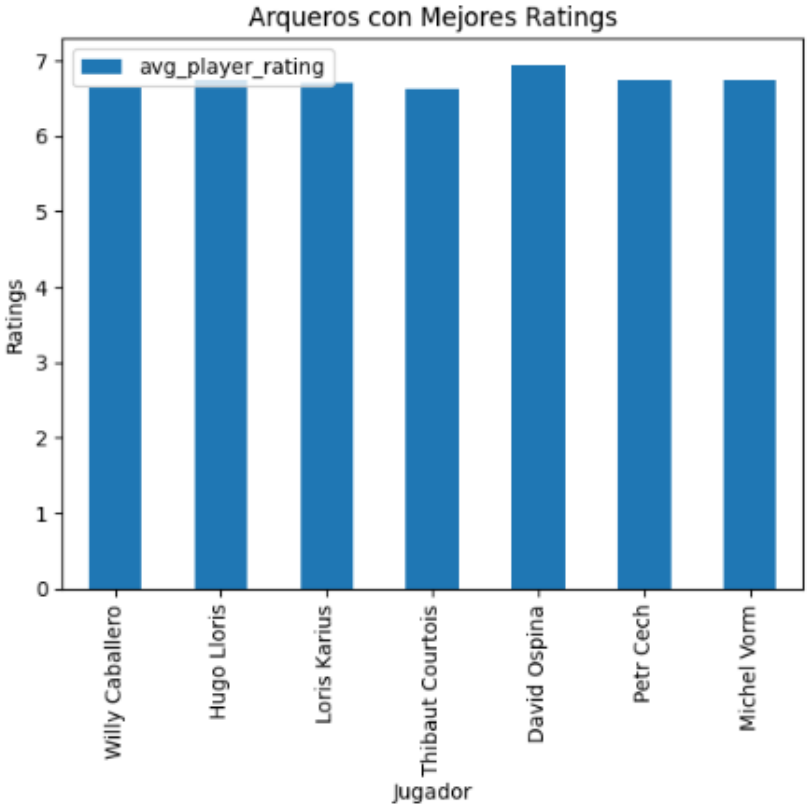
Se crearon gráficas analizando a los jugadores por posición y sus ratings, encontrando así a los mejores candidatos para cada posición. A continuación, se presentan los mejores candidatos de cada posición:



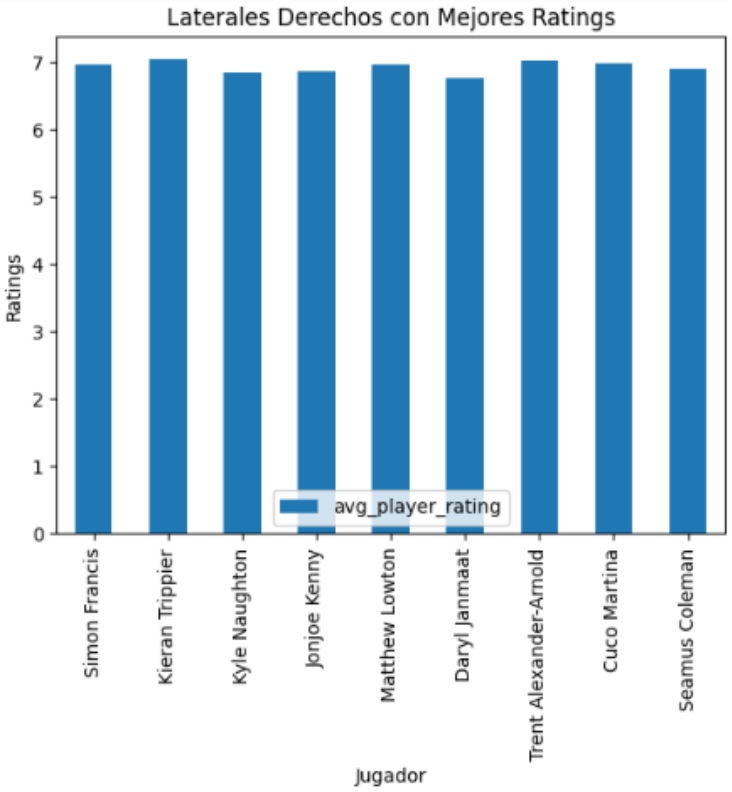
Gráfica 1, mejores jugadores extremo izquierdo



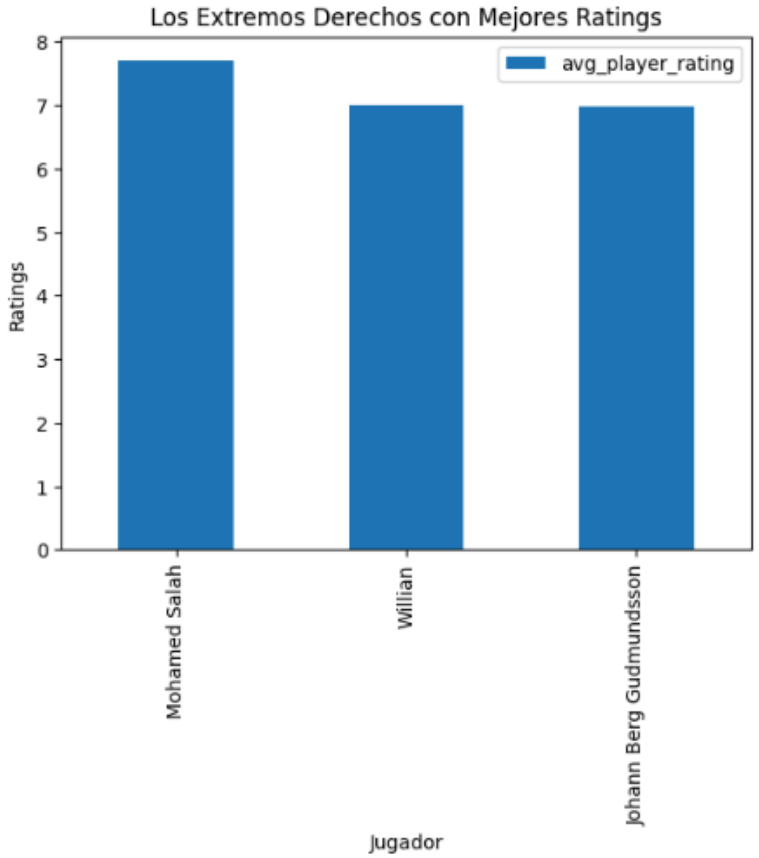
Gráfica 2, mejores jugadores Mediapuntas



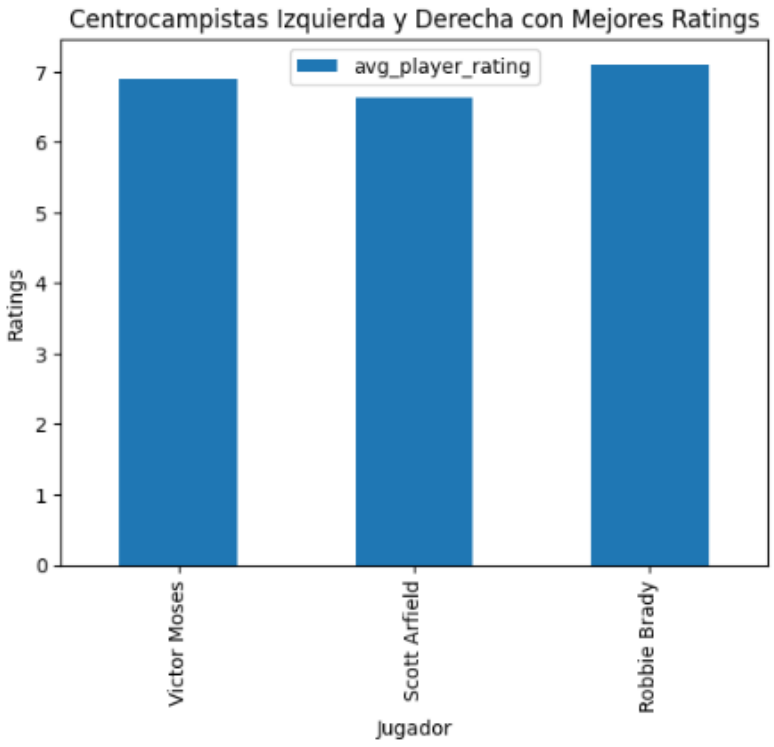
Gráfica 3, mejores jugadores Arqueros



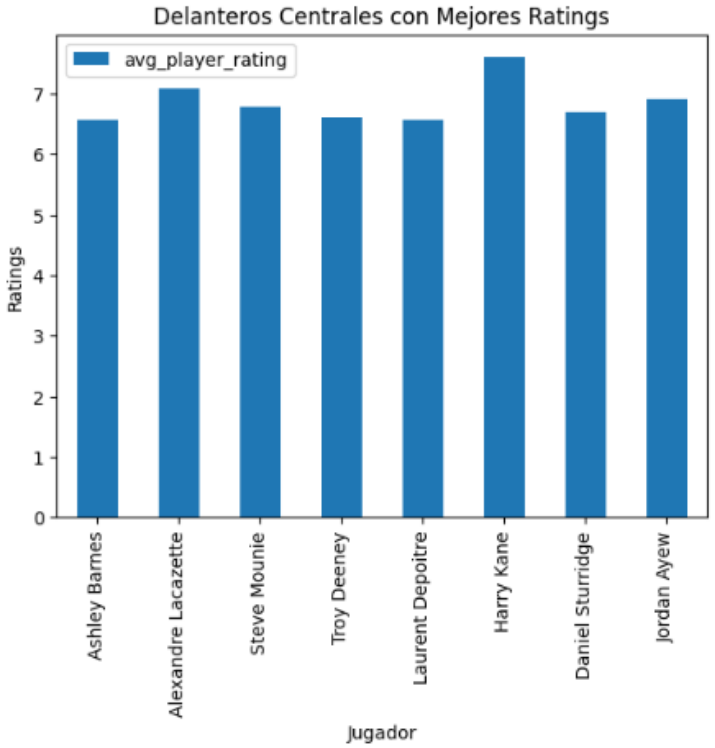
Gráfica 4, mejores jugadores Laterales Derechos



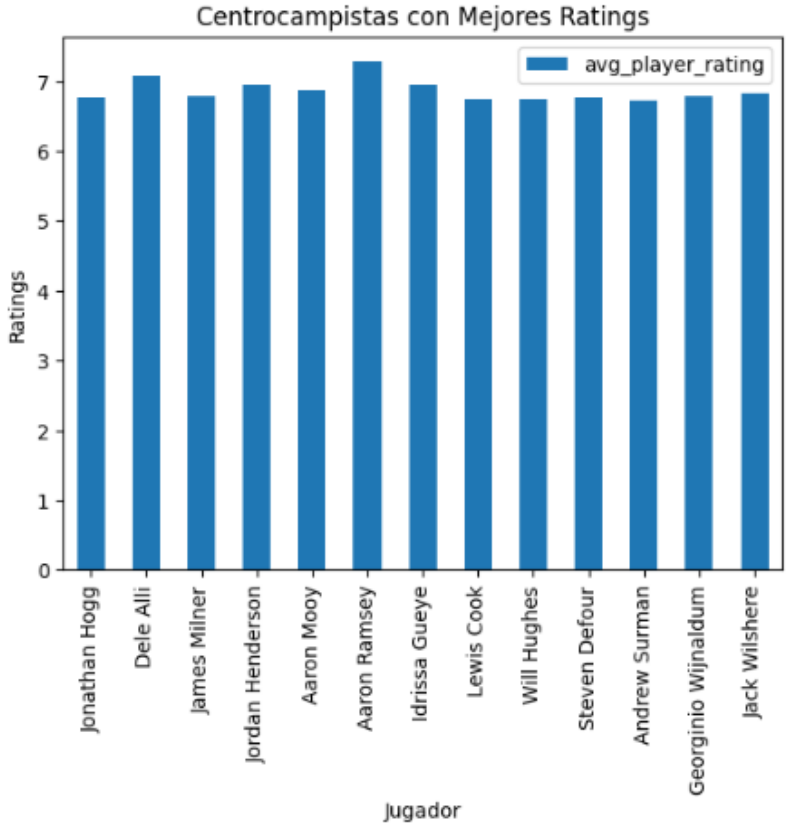
Gráfica 5, mejores jugadores Extremos derechos



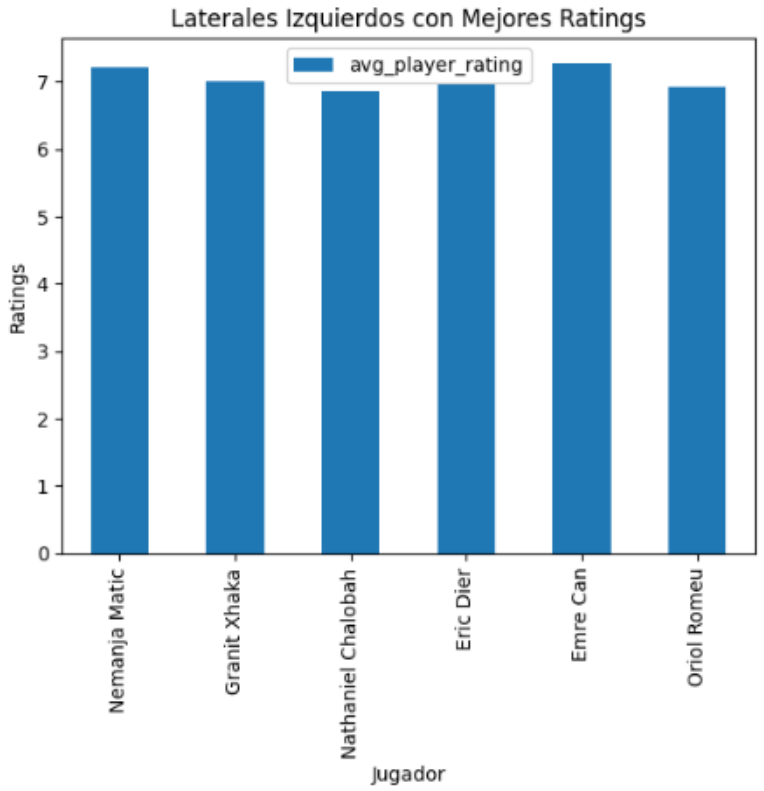
Gráfica 6, mejores jugadores Centrocampistas izquierda/derecha



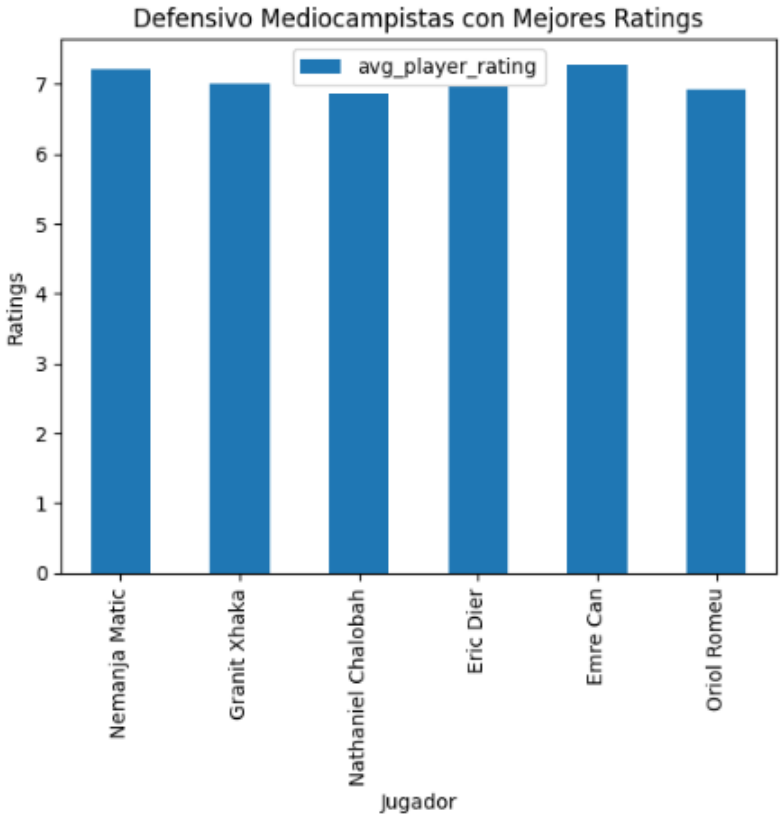
Gráfica 7, mejores jugadores Delanteros Centrales



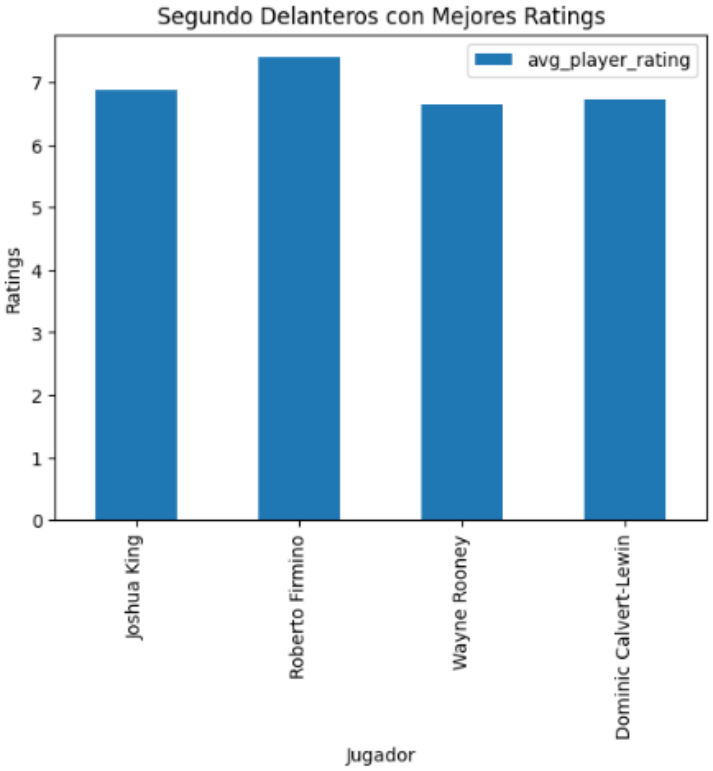
Gráfica 8, mejores jugadores Centrocampistas



Gráfica 9, mejores jugadores Laterales Izquierdos

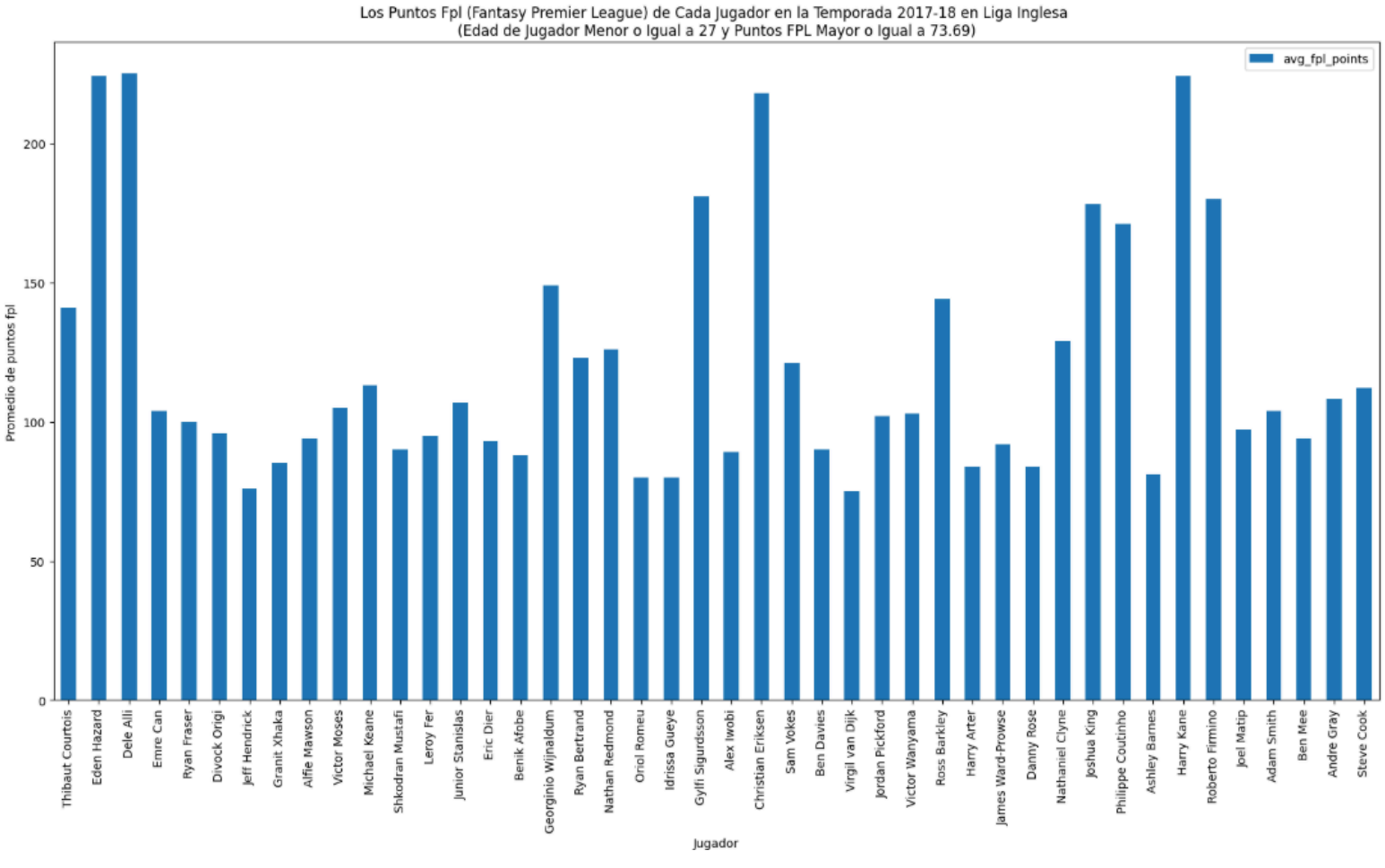


Gráfica 10, mejores jugadores Defensivos Mediocampistas

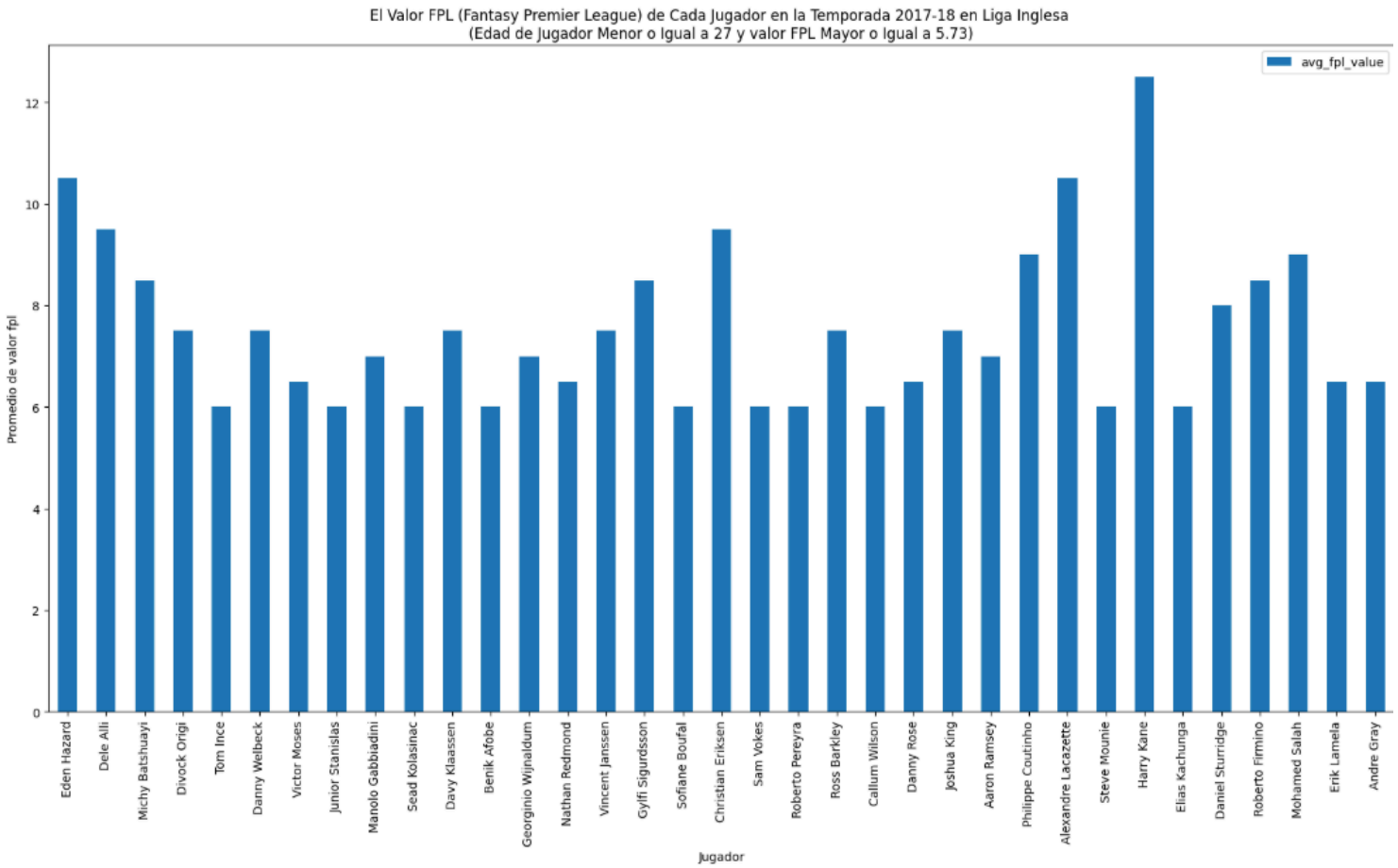


Gráfica 11, mejores jugadores Delanteros

Luego se creó una gráfica general tomando en cuenta valores de la FPL, como valores y points, con tal de obtener la mejor calidad de jugadores según la Fantasy Premier League y sus estadísticas, para así tener un entorno más completo y amplio para la elección.

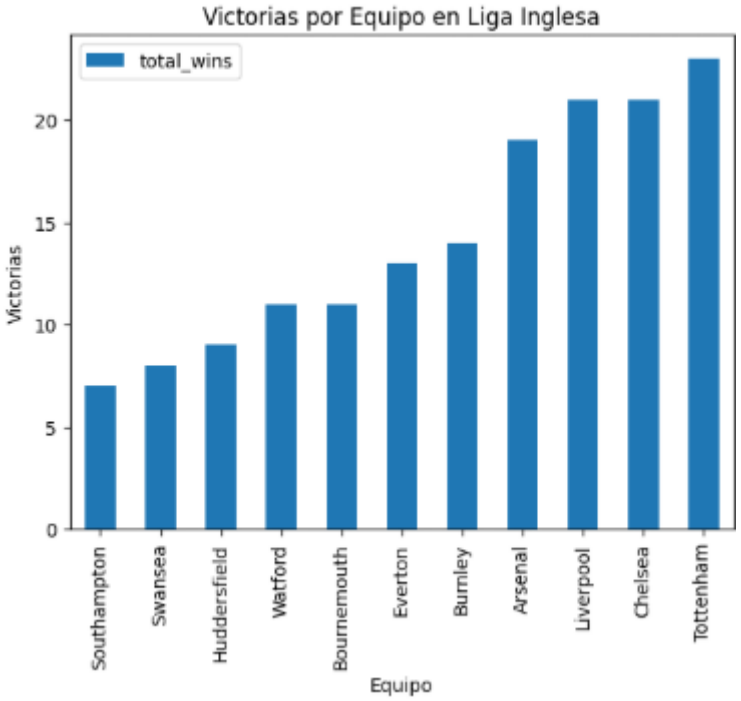


Gráfica 12, mejores jugadores basados en puntaje FPL
(Valores que superan el promedio de 73.69 y menores de 28 años)



Gráfica 13, mejores jugadores basados en valor de jugador FPL
(Valores que superan el promedio de 5.73 y menores de 28 años)

Además de estas gráficas, se consideró analizar a los equipos en sí, por lo tanto, se optó por analizar a los equipos mediante su cantidad de victorias a través de la temporada, creando así la gráfica a continuación:



Gráfica 14, Ganancias por equipo durante la temporada de la liga

Estas gráficas presentan un grupo selecto de jugadores para suplir al nuevo equipo con la mayor calidad, ofreciendo un valor bastante resumido y filtrado de jugadores a elegir, siendo estos quienes destacan frente al resto de jugadores en su campo específico, como por ejemplo Roberto Firmino como segundo delantero o Aaron Ramsey como centrocampista. Por igual es recomendable verificar y analizar a los jugadores que se encuentran en los valores más altos de la FPL, pues estos muchas veces caben dentro de las posiciones ya descritas y pueden mostrar datos de alta calidad y bastante útiles para futuras elecciones, como lo son jugadores como Harry Kane o Eden Hazard. Finalmente, se recomienda elegir candidatos de los equipos de mayor cantidad de ganancias, como lo es primordialmente Tottenham o Chelsea

Finalmente se desea mostrar al equipo más favorable según las estadísticas.
El equipo estaría conformado por (y suponiendo que se necesita uno de cada posición):

1. Eden Hazard - Extremo Izquierdo
2. Philippe Coutinho - Mediapunta
3. David Ospina - Arquero
4. Kieran Trippier - Lateral derecho
5. Mohamed Salah - Extremo derecho
6. Robbie Brady - Centrocampista izquierdo
7. Victor Moses - Centrocampista derecho
8. Harry Kane - Delantero central

9. Aaron Ramsey - Centrocampista
10. Emre Can - Lateral Izquierdo
11. Nemanja Matic - Defensivo Mediocampista
12. Roberto Fimino - Delantero

Estos jugadores se tomaron de sus posiciones mejor evaluadas, sin embargo, todas las demás opciones son equilibradas en comparación y pueden ser consideradas en caso de necesitar otras opciones.

Hallazgos interesantes

Es interesante el encontrar que de tantos jugadores que se mostraron y quienes jugaron a lo largo de estas temporadas, tan pocos sean los que superan al promedio de calificaciones, dado que por cada posición, se encuentran pocos jugadores que destacan frente a los demás, resultando en unos pocos, usualmente menos de una decena, de jugadores que se pueden considerar como los más eficientes a la hora de suplir con su rol en el equipo en el que participen

Este hallazgo fue increíblemente útil a la hora de seleccionar jugadores recomendables para el equipo final, puesto que ofrece un campo de elección más filtrado que permite analizar las estadísticas solo de quienes son considerados como mejores en cada posición o ámbito.

Conclusiones

Una vez realizados los análisis correspondientes, se puede concluir lo siguiente:

De los muchos jugadores a analizar de la liga 17-18 realmente existen pocos que suplan la mayor calidad posible para cada rol que se requiera para el nuevo equipo. En general se optó por crear un equipo de la más alta calidad según ratings y estadísticas de cada jugador en particular. Sin embargo si se desease por nuestros contratistas el ver otras opciones, nuestros valores indican múltiples candidatos a cumplir con el rol deseado, siempre ofreciendo a los mejores jugadores posibles.

En general se optó por crear un equipo con el mejor o mejores valores de los filtros ya generados, produciendo lo que en si debería de ser el equipo ideal a formar para nuestros contratistas.

Resumen

El siguiente documento de parcial aborda un problema de análisis de datos a gran escala en el contexto de equipos de fútbol, tomando en cuenta jugadores, partidos y temporadas para encontrar los mejores candidatos para la creación de un nuevo equipo.

Se tomó en cuenta información de tres archivos:

- Resultados por partidos: conjunto de datos “/csv/resultados_futbol.csv”.
- Estadísticas por equipo: conjunto de datos “/json/temporadas.json”.
- Estadística de jugadores: conjunto de datos “/csv/jugadores.csv”

Utilizando funciones y operaciones para sql en el lenguaje de programación Python con tal de obtener resultados de análisis eficientes y efectivos para conclusiones y resultados concisos.