Procesamiento de Datos

Juan Manuel Soto

PONTIFICIA UNIVERSIDAD JAVERIANA
Facultad de Ciencias

Semestre 23-03

INDICE:

1. Introducción:

En este documento se plasman diferentes prácticas para resolver un problema de toma de decisiones. Para este problema, se usó un servicio para hacer las diferentes consultas llamadas DataBricks en un framework que se llama Spark, haciendo visualizaciones para entender mejor los resultados, y observar con una mejor claridad qué datos son los más importantes que debemos usar para el análisis del problema

2. Contexto general del problema

el problema surge a partir de una suposición de que somos contactados por un inversor millonario que quiere crear un nuevo equipo de fútbol en la liga inglesa y tiene el interés de conocer un poco más los equipos y resultados que se obtuvieron en la liga de 17-18, para tener una mejor idea de que tipos de jugadores preferiría contratar y que estilo de juego quiere que tenga su equipo. A continuación se observa que bases de datos se van a utilizar junto con una breve descripción:

- Resultados por partidos: conjunto de datos “results.csv”.
- Estadísticas por equipo: conjunto de datos “season_stats.json”
- Estadística de jugadores: conjunto de datos “players.csv”

El fin de este problema es buscar que jugadores tengan mejores resultados según sus descripciones y tener los mejores resultados en los equipos que están jugando.

3. Descripción de los conjuntos de datos y sus variables más importantes y Descriptivas generales de las variables

las bases que datos que se usarán en este caso serán:

- Resultados por partidos: conjunto de datos “results.csv”
- Estadística de jugadores: conjunto de datos “players.csv”

ya que tienen tipos de datos más importantes para poder resolver el problema planteado

3.1 Base de datos “players.csv”

la base de datos está conformada por los siguientes atributos o valores:

- 'name'= nombre
- 'club' = equipo
- 'age'= edad
- 'position'= posicion
- 'position_cat'= posicion alterna
- 'market_value'= valor de mercado
- 'page_views'= paginas vistas
- 'fpl_value' =valor fpl
- 'fpl_sel' =sel fpl
- 'fpl_points"= puntos de fpl
- 'region'= region
- 'nationality'= nacionalidad
- 'new_foreign' = nuevo extranjero
- 'age_cat' = edad cat
- 'club_id" = id del club
- 'big_club' = club grande
- 'new_signing'= nueva forma

de los cuales los datos más importantes para poder solucionar el problema son:

- 'club'
- 'position'
- 'fpl_points'

3.2 Base de datos “results.csv”

la base de datos está conformada por los siguientes atributos o valores:

- 'Season'= temporada
- 'DateTime'= dia y hora
- 'HomeTeam'= equipo local
- 'AwayTeam'= equipo visitante
- 'HTR'= referencia a si gano, empato y perdio el equipo

descripciones, como faltas, tarjetas rojas, amarillas, tiros de arco etc:

- 'FTHG'
- 'FTAG'
- 'FTR'
- 'HTHG'
- 'HTAG'
- 'Referee'
- 'HS'
- 'AS'
- 'HST'
- 'AST'
- 'HC'
- 'AC'
- 'HF'
- 'AF'
- 'HY'
- 'AY'
- 'HR'
- 'AR'

de los cuales los datos más importantes para poder solucionar el problema son:

- 'HTR'
- 'HomeTeam'
- 'AwayTeam'
- 'Season'

4. Calidad de datos de los conjuntos de datos

la base de datos “results.csv” tenía valores vacíos y valores que no nos interesaba ya que no ayuda a resolver el problema, entonces se llevó a cabo lo siguientes pasos:

1. se abrió un Notebook en Colab
2. se leyeron los archivos csv
3. se observó que valores nulos tenias
4. se eliminan los valores nulos o vacíos
5. se vuelve a rectificar la base de datos
6. se descarga la base de datos ya limpia

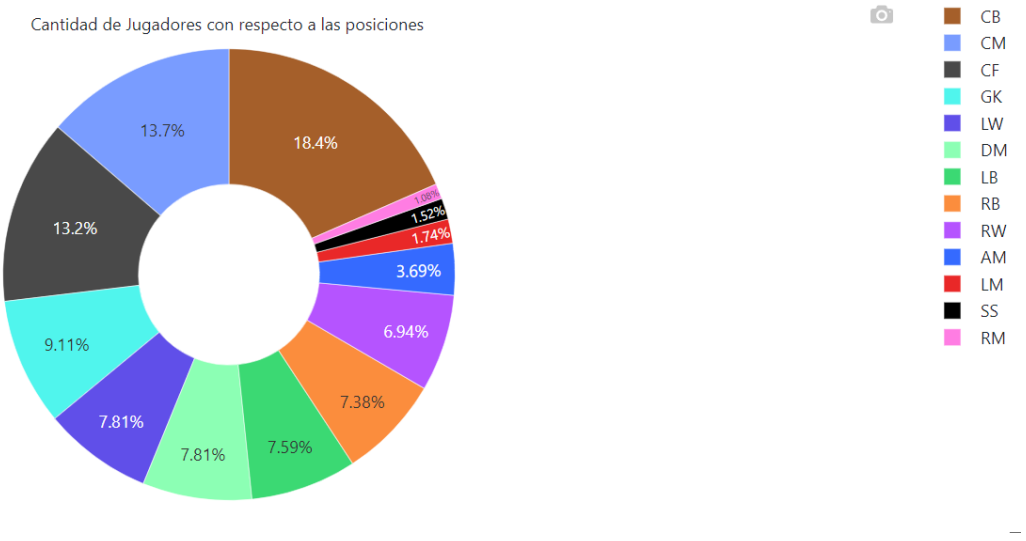
el proceso se puede observar a través del siguiente link:

https://colab.research.google.com/drive/1sl5h4r3VjVa0wC6BR84kRNaDpP6ynBwZ?usp=sharing

5. Hallazgos interesantes del problema de negocio

unas cosas interesantes durante el proceso y desarrollo del problema fueron los siguientes:

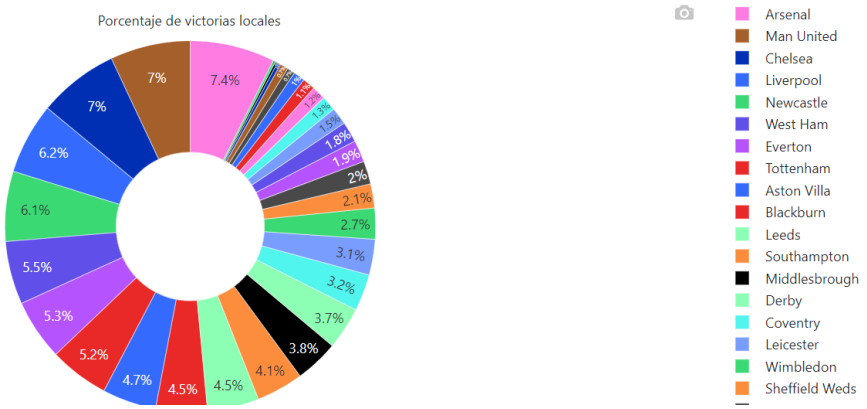
1. dentro de la base de datos de los jugadores hay una gran cantidad de esquinero(CB) como lo muestra la siguiente gráfica siendo la posición CB que predomina en los jugadores



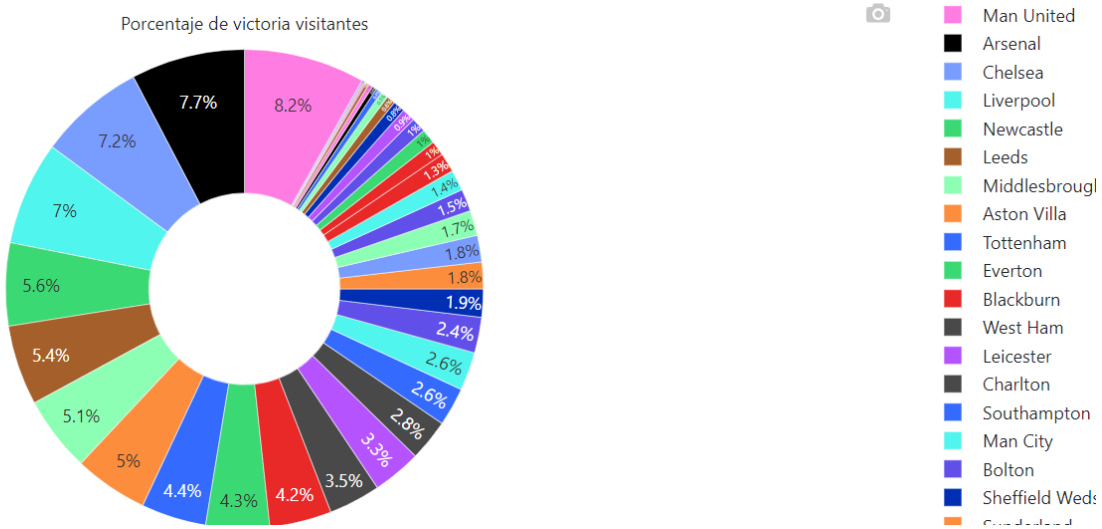
2. las estadísticas de los partidos en la siguiente imagen muestra datos (“mencionados en el apartado Descripción de los conjuntos”) vacios, suponiendo que los partidos fueron en constante movimiento sin haber ningún corte de tiempo en cada uno de ellos

HTHG	HTAG	HTR	Referee	HS	AS	HST	AST	HC
NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA

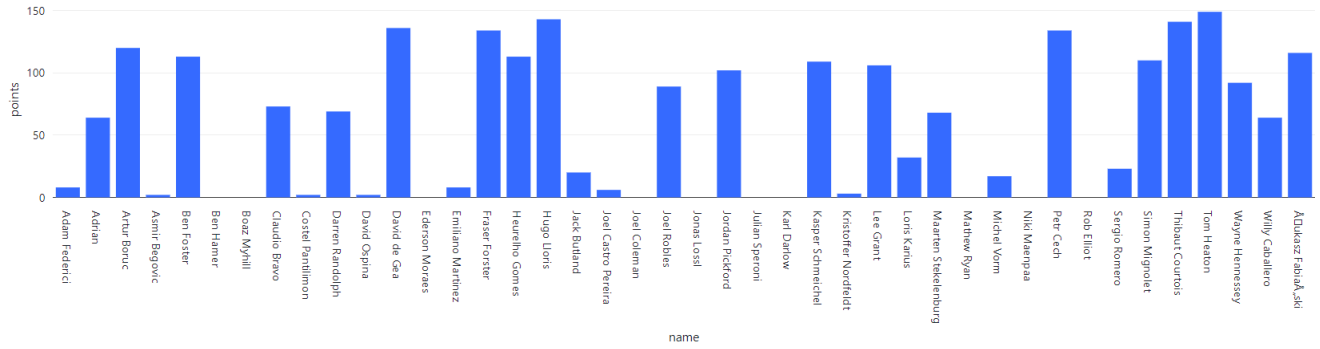
3. en la siguiente imagen, podemos observar que el índice de victorias jugando de equipo local, lo tiene Arsenal, Manchester United y Chelsea, teniendo una tasa de victorias muy alta dentro de todos los equipos de la base de datos



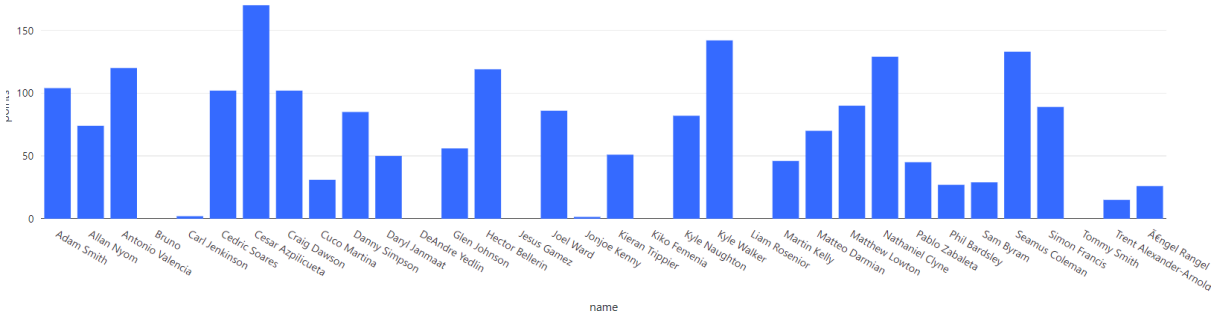
4.para el caso de los equipos visitantes podemos ver que también tienen una alta tasa de victorias los anteriores equipos siendo Manchester United, Arsenal y Chelsea



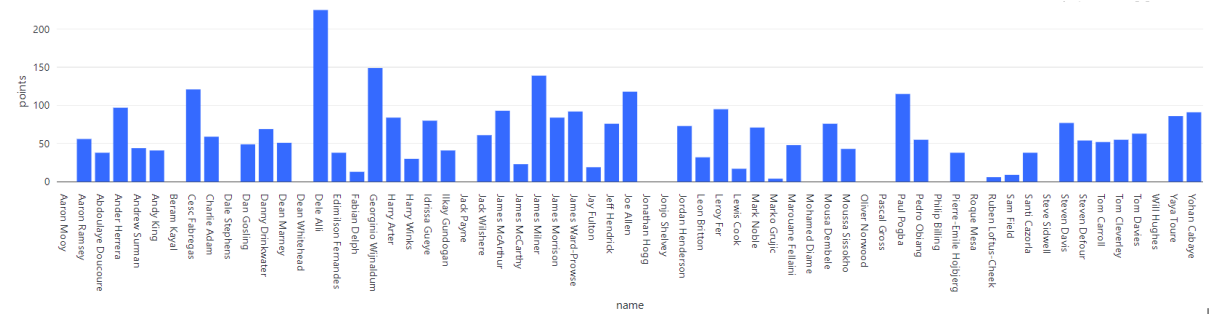
5. las en estadísticas de los porteros, mencionan que Tom Heaton es el que mejor desempeño tiene



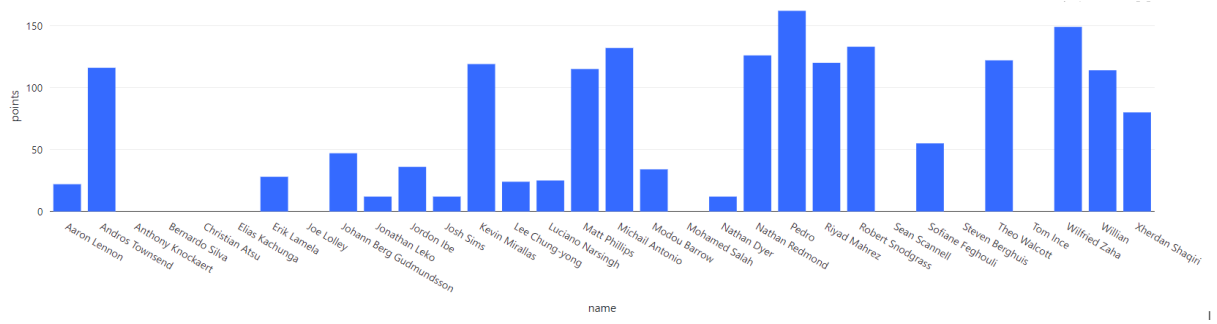
6.las estadísticas de los laterales derechos ,cesar Azpilicueta es el que mejor desempeño tiene



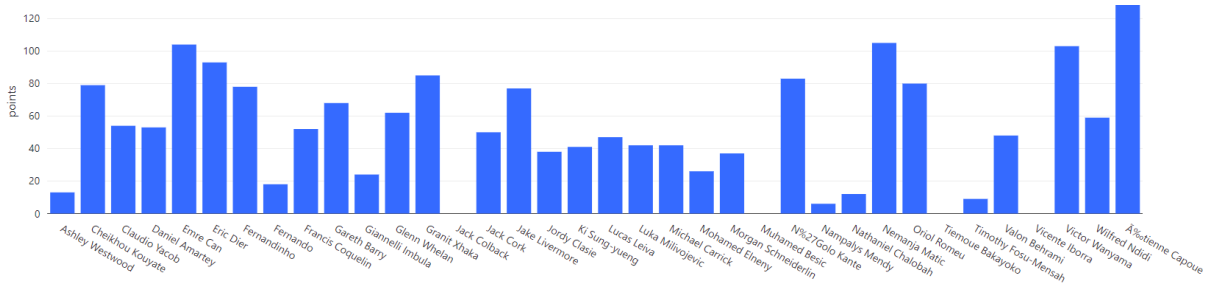
7.las estadísticas del medio centro,Dele Alli es el que mejor desempeño tiene



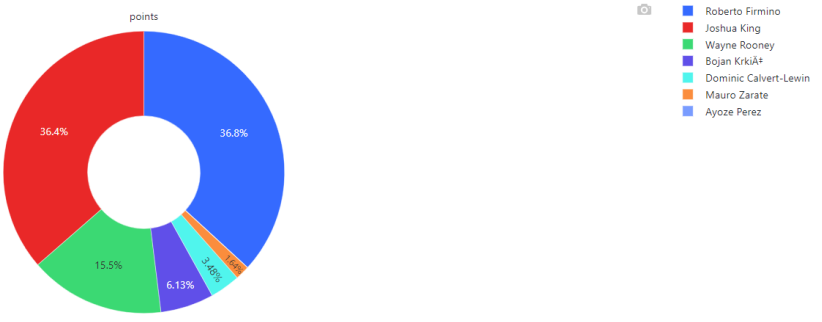
8.las estadísticas de los extremos izquierdos,Pedro es el que mejor desempeño tiene



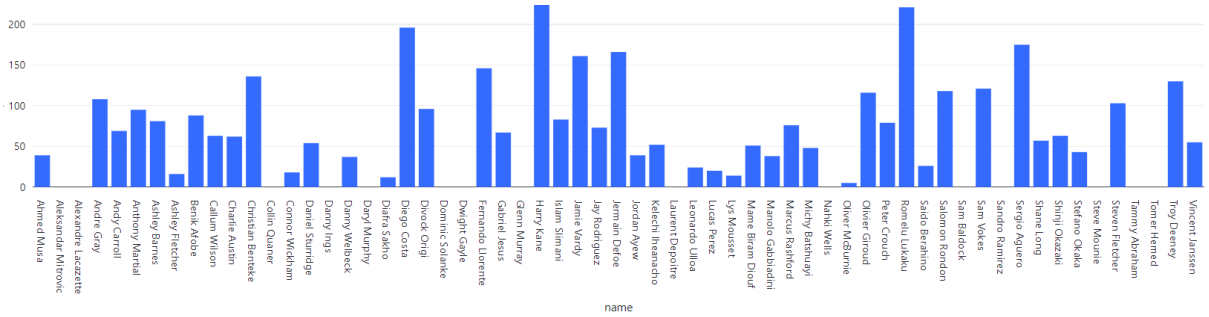
9.las estadísticas de los mediocampistas defensivo diestro ,Capoue es el que mejor desempeño tiene



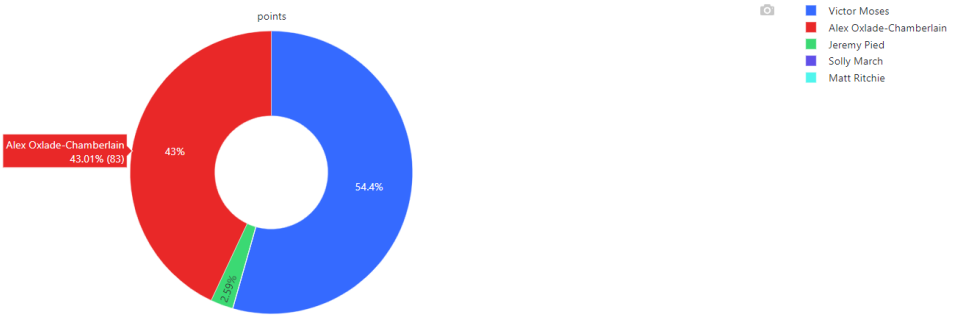
10.las estadísticas del strong safery ,roberto Firmino es el que mejor desempeño tiene



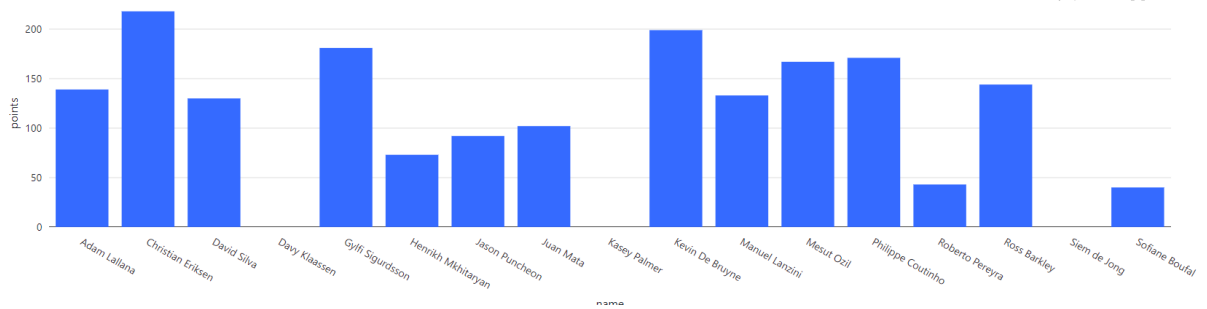
11.las estadísticas del segundo delantero derech ,Harry Kane es el que mejor desempeño tiene



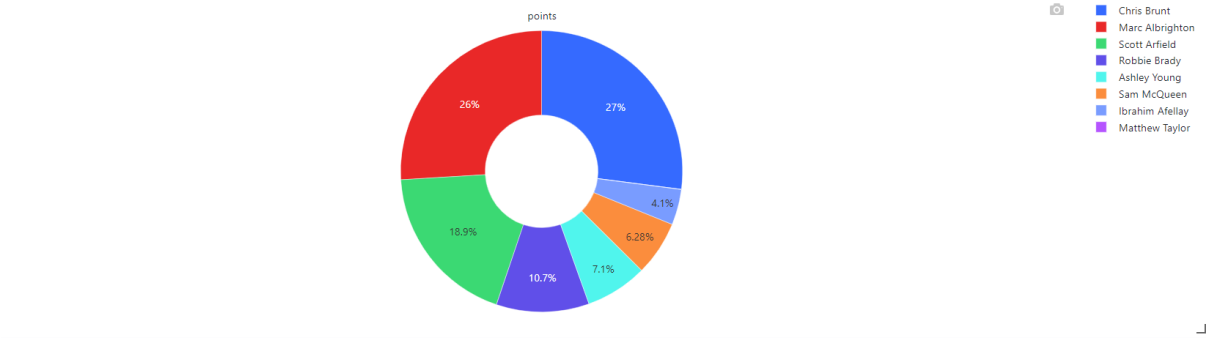
12.las estadísticas del medio derecho ,Victor Moses es el que mejor desempeño tiene



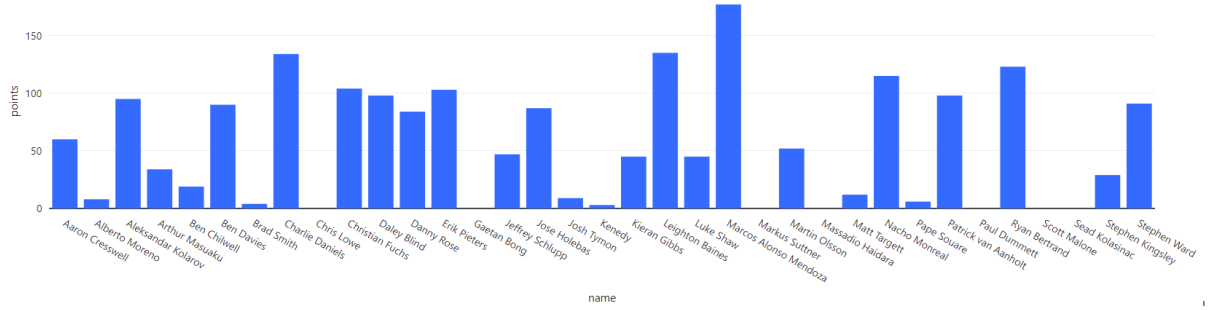
13.las estadísticas del volante ,Christian Eriksen es el que mejor desempeño tiene



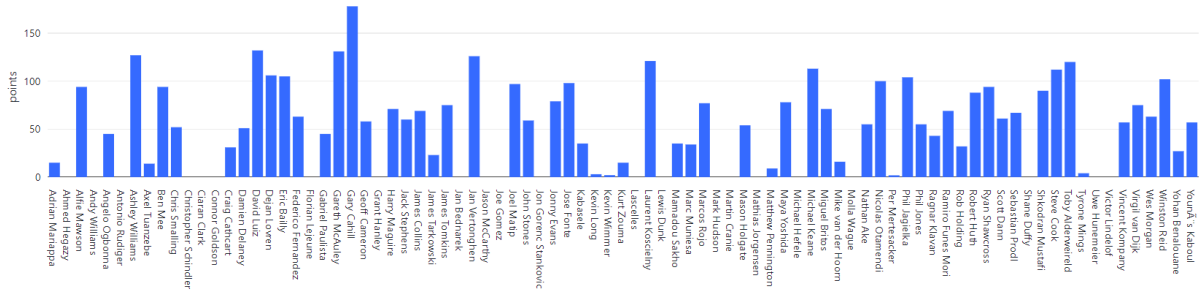
14.las estadísticas del medio izquierdo,Chirs Brunt es el que mejor desempeño tiene



15.las estadísticas del lateral izquierdo,Marcos Mendoza Firmino es el que mejor desempeño tiene



16 las estadísticas de los esquineros ,Gary Cahill es el que mejor desempeño tiene



Conclusiones

los mejores jugadores para cada posición del equipo están notadas en los gráficos anteriores de los hallazgos, y como resultado podemos decirle al inversor que lo mejor que puede invertir seria en los 3 equipos que obtuvieron los mejores resultados o porcentajes siendo Arsenal, Manchester United y Chelsea los mejores equipos en los que podemos elegir buenos jugadores para que el inversor cree el equipo.

por otra Parte, podemos decirle al inversor que el equipo puede tener un estilo de juego atacante o estándar , ya que los resultados son más altos y favorables en las posiciones de los delanteros, volantes y centros, por lo cual puede tomar el equipo un estilo más de atacante o puede tomar un estilo atacante-defensivo ya que los resultados también de las defensas son muy buenas