Análisis Avanzado de Jugadores en la Bundesliga Utilizando PCA y Clustering K-Means

Contacto

Emanuel Martín Novelo Hernández

emnovelo98@gmail.com

https://github.com/EmanuelNovelo/Bundesliga-Player

Descripción y Objetivo

Objetivo del proyecto: Presentación y reporte técnico de resultados donde se aplique una o varias técnicas multivariadas en el análisis de datos de varias variables aplicado a la solución de un problema real

Este proyecto se centra en el análisis avanzado de jugadores en la Bundesliga utilizando técnicas de PCA (Análisis de Componentes Principales) y Clustering K-Means.

Objetivo específico: El objetivo principal es segmentar a los jugadores de la Bundesliga en grupos basados en sus características y desempeño en el campo.

Introducción

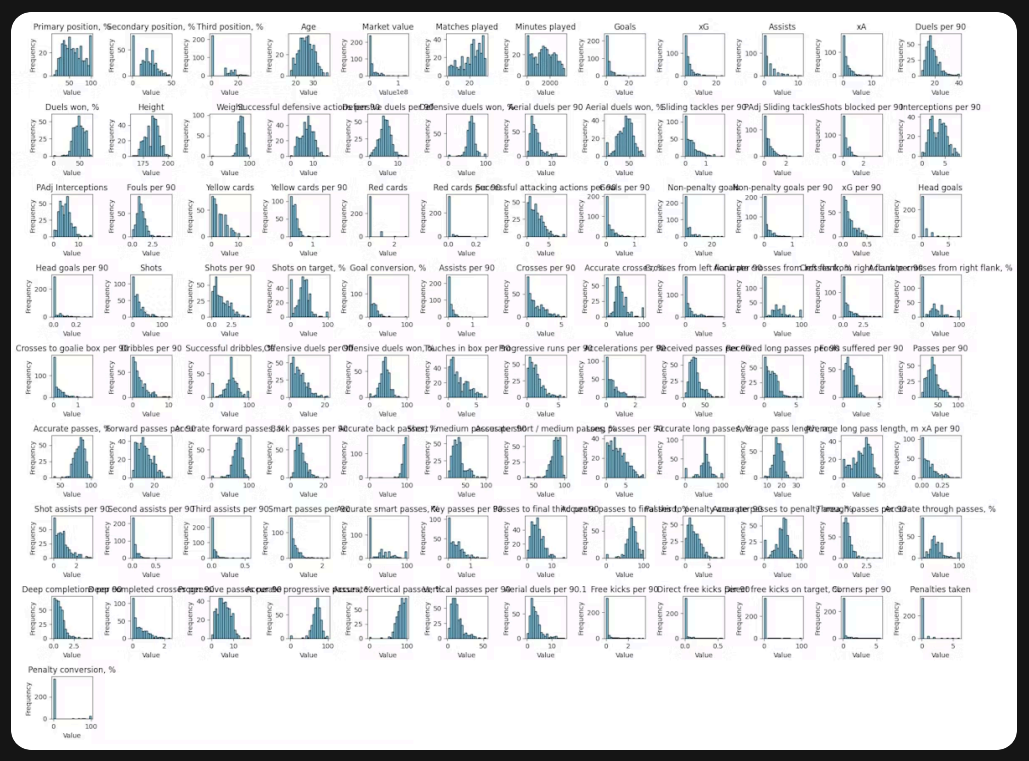
El conjunto de datos utilizado contiene estadísticas de jugadores y sus posiciones en el campo. Cabe destacar que los porteros fueron excluidos del análisis debido a la naturaleza única de su posición.

Se realizó también un análisis exploratorio de Datos en donde se encontraron valores faltantes. Estos fueron reemplazados con la media, o eliminados si sus valores faltantes excedian el 50% de la información.

	Player	Full name	Wyscout id	Team	Team within selected timeframe	Team logo	Competition	Position	Primary position	Primary position, %	...	Accurate progressive passes, %	Accurate vertical passes, %	Vertical passes per 90	Aerial duels per 90.1	Free kicks per 90	Direct free kicks per 90	Direct free kicks on target, %	Corners per 90	Penalties taken	co
4	W. Pacho	Willian Joel Pacho Tenorio	-74304	Eintracht Frankfurt	Eintracht Frankfurt	https://cdn5.wyscout.com/photos/team/public/9_...	Bundesliga	LCB3, LCB	LCB3	60	...	59.16	95.91	24.74	3.50	0.00	0.0	0.0	0.0	0	
6	P. Mainka	Patrick Mainka	260049	Heidenheim	Heidenheim	https://cdn5.wyscout.com/photos/team/public/37...	Bundesliga	RCB	RCB	98	...	72.44	96.80	14.67	5.88	0.03	0.0	0.0	0.0	0	
8	Bernardo	Bernardo Fernandes da Silva Junior	328528	Bochum	Bochum	https://cdn5.wyscout.com/photos/team/public/97...	Bundesliga	LB, LCB3	LB	63	...	70.82	88.03	8.81	8.04	0.00	0.0	0.0	0.0	0	
10	F. Uduokhai	Felix Uduokhai	359831	Augsburg	Augsburg	https://cdn5.wyscout.com/photos/team/public/34...	Bundesliga	LCB, LCB3	LCB	76	...	68.63	94.01	18.71	5.69	0.00	0.0	0.0	0.0	0	
11	W. Anton	Waldemar Anton	307802	Stuttgart	Stuttgart	https://cdn5.wyscout.com/photos/team/public/96...	Bundesliga	RCB, CB, LCB	RCB	56	...	76.32	96.45	31.60	4.35	0.00	0.0	0.0	0.0	0	

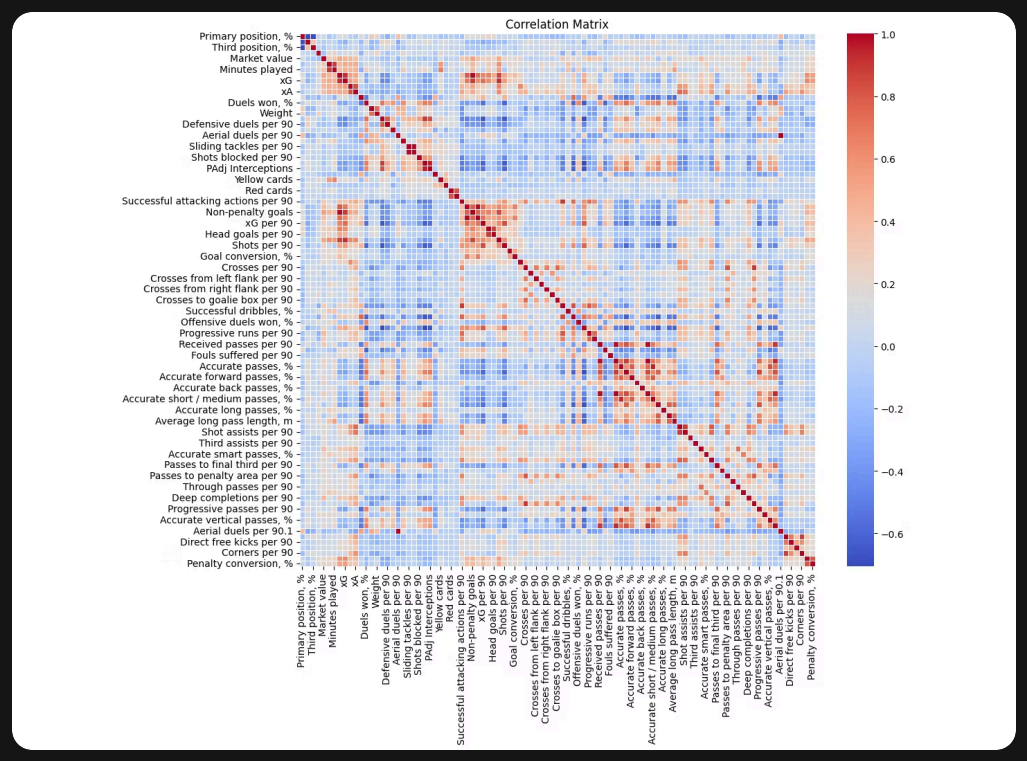
5 rows × 114 columns

Descripción de las Variables y Normalidad Univariada



Análisis de Histogramas

Durante el análisis, se examinaron los histogramas de las variables para evaluar su distribución.



Matriz de Correlación

El volumen de variables era alto, por lo que un mapa de “calor” ayuda a identificar correlaciones fuertes tanto positiva como negativamente

Normalidad Univariada

Solo la variable 'Defensive duels per 90' proviene de una distribución normal

```
Variable 'Defensive duels per 90' satisfies the normality assumption.
Shapiro-Wilk test statistic: 0.9928457736968994
p-value: 0.060401950031518936
------------------------------------------------
```

Se muestra el vector de medias muestrales para algunas variables

	Primary position, %	Secondary position, %	Third position, %	Age	Market value	Matches played	Minutes played	Goals	xG	Assists	...	Accurate progressive passes, %	Accurate vertical passes, %	Vertical passes per 90	Aerial duels per 90.1	Free kicks per 90
0	1.278612e-16	-1.369941e-17	-3.424853e-17	-1.826588e-16	3.653176e-17	0.0	-7.306352e-17	-7.306352e-17	-1.095953e-16	3.653176e-17	...	1.004623e-16	9.589587e-16	1.095953e-16	-1.095953e-16	4.109823e-17

Se muestra el vector de desviaciones muestrales para algunas variables

	Primary position, %	Secondary position, %	Third position, %	Age	Market value	Matches played	Minutes played	Goals	xG	Assists	...	Accurate progressive passes, %	Accurate vertical passes, %	Vertical passes per 90	Aerial duels per 90.1	Free kicks per 90	Direct free kicks per 90	Direct free kicks on target, %
0	1.001288	1.001288	1.001288	1.001288	1.001288	1.001288	1.001288	1.001288	1.001288	1.001288	...	1.001288	1.001288	1.001288	1.001288	1.001288	1.001288	1.001288

Normalidad Multivariada

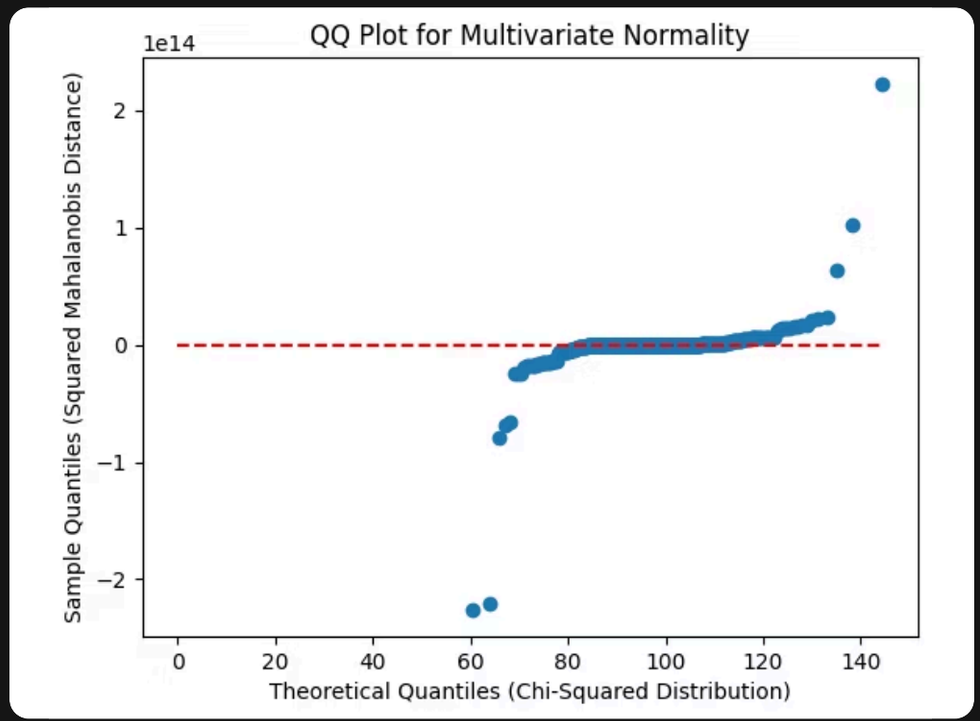
Test de Henze-Zirkler

El test de Henze-Zirkler se utiliza para evaluar la normalidad multivariada de los datos. En nuestro análisis de jugadores en la Bundesliga, encontramos que los datos no siguen una distribución normal multivariada.

```
HZResults(hz=1556, pval=0.0, normal=False)
```

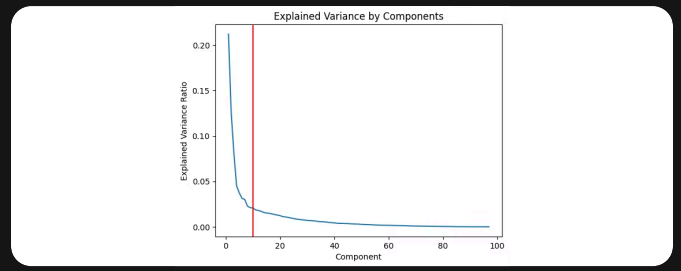
Gráficos QQ

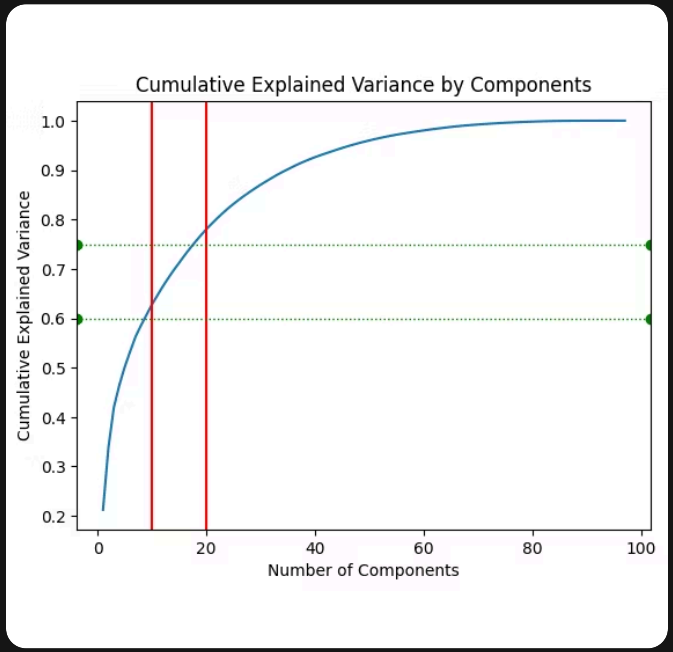
En nuestro análisis, los gráficos QQ mostraron desviaciones significativas de la línea de referencia, lo que indica que los datos no siguen una distribución normal.



Análisis PCA

Se realizo un PCA en donde de los cerca de 100 componentes se eligieron 20, que explican el 78% de la varianza.





En el gráfico de la varianza acumulada se percibe de mejor forma el porcentaje de la varianza explicada en el componente 20.

	Eigenvalue	Explained Variance	Cumulative Variance
Component 1	20.631754	0.212152	0.212152
Component 2	12.275542	0.126227	0.338378
Component 3	7.884430	0.081074	0.419452
Component 4	4.374843	0.044986	0.464438
Component 5	3.615560	0.037178	0.501616
Component 6	3.043227	0.031293	0.532909
Component 7	2.928163	0.030110	0.563018
Component 8	2.214568	0.022772	0.585790
Component 9	2.047307	0.021052	0.606842
Component 10	2.024297	0.020815	0.627658
Component 11	1.808548	0.018597	0.646254
Component 12	1.749539	0.017990	0.664245
Component 13	1.663612	0.017107	0.681351
Component 14	1.533581	0.015769	0.697121
Component 15	1.472971	0.015146	0.712267
Component 16	1.444769	0.014856	0.727123
Component 17	1.390209	0.014295	0.741418
Component 18	1.316126	0.013533	0.754952
Component 19	1.258165	0.012937	0.767889
Component 20	1.195564	0.012294	0.780183

Puntuaciones

En el contexto del análisis de jugadores en la Bundesliga, el PCA nos permite identificar las variables más importantes que contribuyen a las diferencias entre los jugadores y agruparlos en categorías similares.

Principales variables por Componentes

	Component 1	Component 2	Component 3	Component 4	Component 5	Component 6	Component 7	Component 8	Component 9	Component 10	Component 11	Component 12	Component 13	Component 14	Component 15
0	Touches in box per 90	Aerial duels per 90.1	xG	Accurate passes, %	Primary position, %	Free kicks per 90	Accurate crosses, %	Goal conversion, %	Accurate passes to penalty area, %	PAdj Sliding tackles	Red cards	Red cards per 90	Red cards per 90	Accurate crosses from left flank, %	Accurate smart passes, %
1	Forward passes per 90	Aerial duels per 90	Head goals	Short / medium passes per 90	Sliding tackles per 90	Direct free kicks per 90	Through passes per 90	Non-penalty goals per 90	PAdj Sliding tackles	Sliding tackles per 90	Red cards per 90	Red cards	Accurate through passes, %	Accurate through passes, %	Accurate through passes, %
2	Interceptions per 90	xA	Goals	Accurate passes to final third, %	Yellow cards	Corners per 90	PAdj Sliding tackles	Goals per 90	Accurate crosses, %	Defensive duels won, %	Second assists per 90	Second assists per 90	Through passes per 90	Accurate smart passes, %	Accurate crosses, %
3	Successful defensive actions per 90	Passes to penalty area per 90	Non-penalty goals	Received passes per 90	Fouls suffered per 90	Crosses per 90	Accurate passes to penalty area, %	Secondary position, %	Sliding tackles per 90	Third position, %	Shots on target, %	Assists	Red cards	Accurate crosses, %	Fouls suffered per 90
4	Duels won, %	Accurate progressive passes, %	Minutes played	Back passes per 90	PAdj Sliding tackles	Penalties taken	Sliding tackles per 90	Shots on target, %	Direct free kicks per 90	Primary position, %	Goal conversion, %	Minutes played	Smart passes per 90	Third position, %	Accurate crosses from right flank, %
5	Shots per 90	xA per 90	Shots	Age	Fouls per 90	Crosses from right flank per 90	Smart passes per 90	Successful dribbles, %	Accelerations per 90	Accurate passes to final third, %	Corners per 90	xA	Crosses to goalie box per 90	Primary position, %	Accurate passes to penalty area, %
6	PAdj Interceptions	Assists	Head goals per 90	Yellow cards	Defensive duels per 90	Penalty conversion, %	Accurate crosses from right flank, %	Accurate back passes, %	Age	Direct free kicks per 90	Free kicks per 90	Yellow cards	Assists per 90	Head goals per 90	Through passes per 90
7	Offensive duels per 90	Deep completed crosses per 90	Goals per 90	Average pass length, m	Yellow cards per 90	Deep completed crosses per 90	Accurate crosses from left flank, %	Primary position, %	Smart passes per 90	Secondary position, %	Duels per 90	Matches played	Second assists per 90	Red cards	Third position, %
8	xG per 90	Shot assists per 90	Non-penalty goals per 90	Accurate forward passes, %	Dribbles per 90	Crosses to goalie box per 90	Passes to final third per 90	Third position, %	Progressive runs per 90	Fouls suffered per 90	Aerial duels per 90.1	Head goals per 90	Accurate smart passes, %	Red cards per 90	Secondary position, %

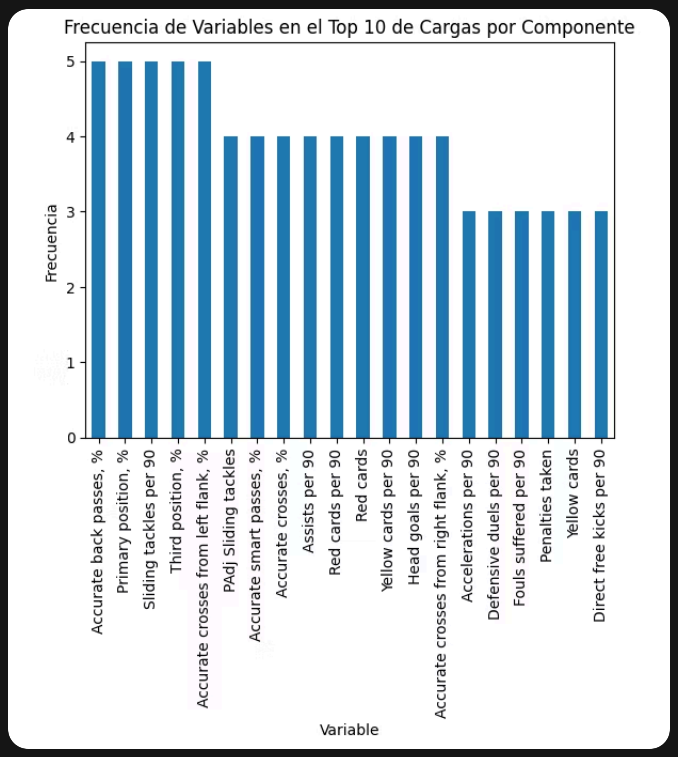
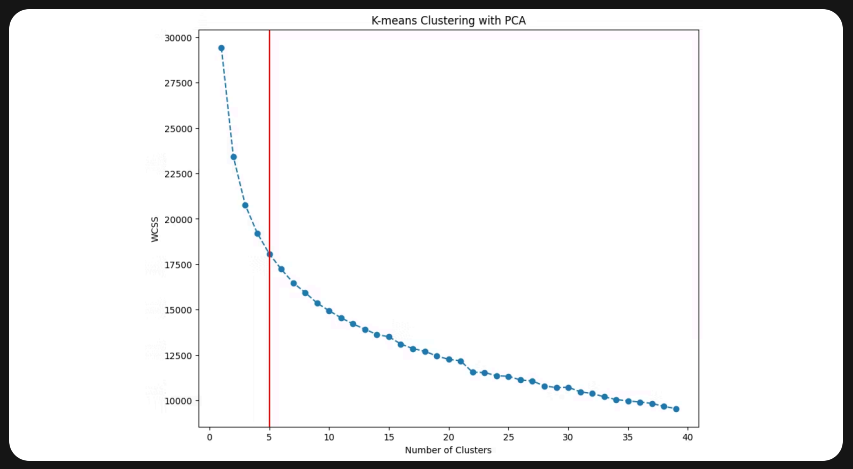


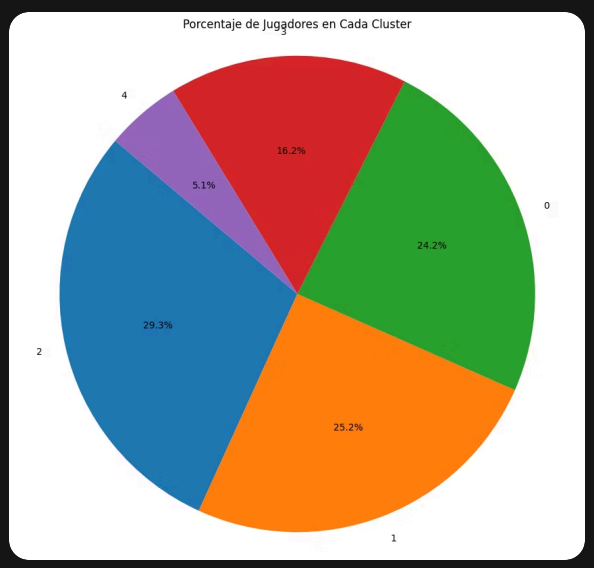
Tabla final - con Componentes (20)

	Player	Full name	Wyscout id	Team	Team within selected timeframe	Team logo	Competition	Position	Primary position	Primary position, %	...	Component 12	Component 13	Component 14	Component 15	Component 16	Component 17
0	W. Pacho	Willian Joel Pacho Tenorio	-74304	Eintracht Frankfurt	Eintracht Frankfurt	https://cdn5.wyscout.com/photos/team/public/9_...	Bundesliga	LCB3, LCB	LCB3	60	...	-1.662673	-0.689761	1.224412	-0.285383	0.857602	-0.144082
1	P. Mainka	Patrick Mainka	260049	Heidenheim	Heidenheim	https://cdn5.wyscout.com/photos/team/public/37...	Bundesliga	RCB	RCB	98	...	-0.868211	-0.691994	-1.144964	-1.850234	-0.253425	0.530775
2	Bernardo	Bernardo Fernandes da Silva Junior	328528	Bochum	Bochum	https://cdn5.wyscout.com/photos/team/public/97...	Bundesliga	LB, LCB3	LB	63	...	-1.816966	-0.276582	-0.097691	1.162612	-0.618241	-0.875638
3	F. Uduokhai	Felix Uduokhai	359831	Augsburg	Augsburg	https://cdn5.wyscout.com/photos/team/public/34...	Bundesliga	LCB, LCB3	LCB	76	...	-0.124731	1.061685	0.916057	-0.237660	0.414548	0.311141
4	W. Anton	Waldemar Anton	307802	Stuttgart	Stuttgart	https://cdn5.wyscout.com/photos/team/public/96...	Bundesliga	RCB, CB, LCB	RCB	56	...	-1.659126	0.075131	-0.073524	0.746975	0.859127	-1.068669

Clustering K-Means & Conclusiones

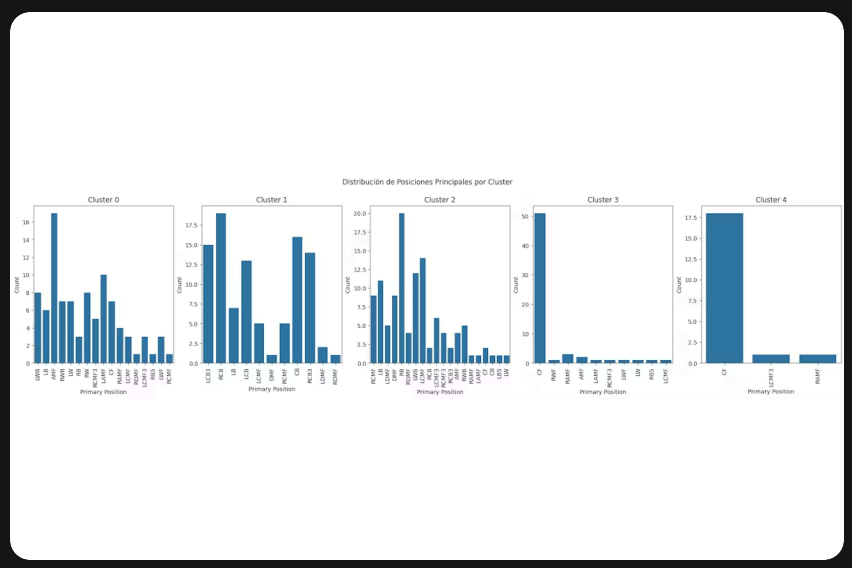


Se realizó un análisis previo para definir el número de Clusters, se gráfico la suma de las varianzas (wcss) del método de K-Means vs un numéro máximo arbitriario de clusters (40 en este caso) para definir el número idea de clusters



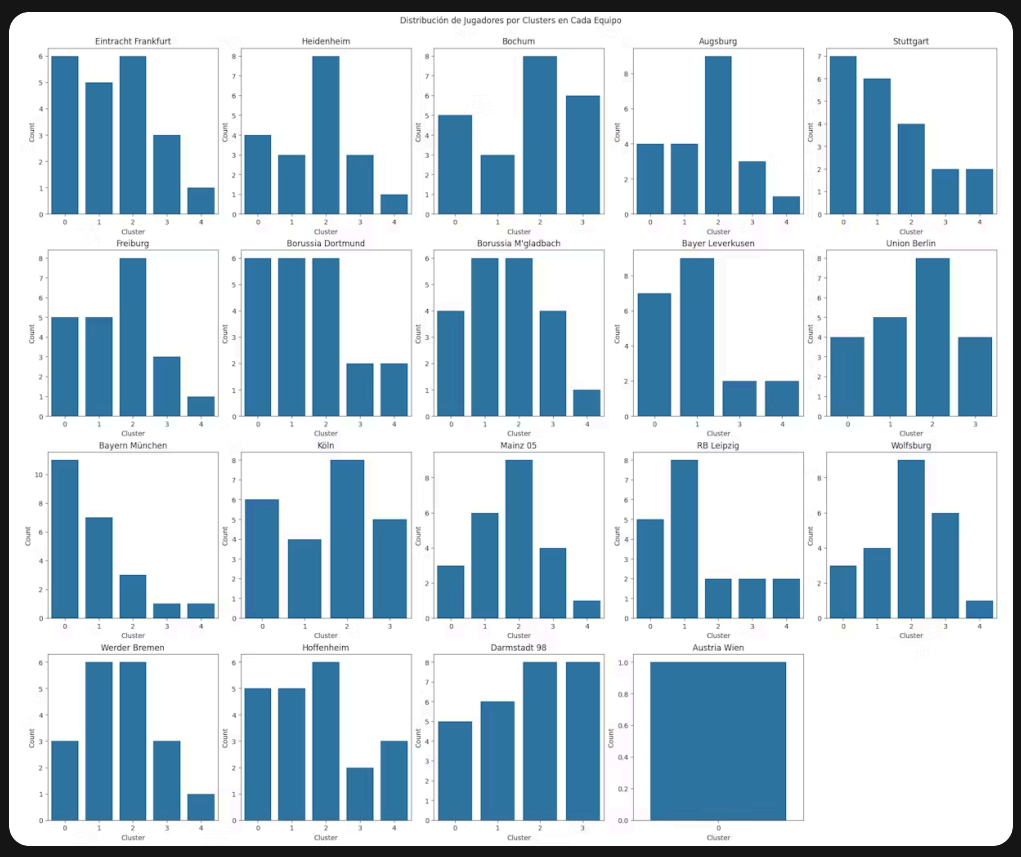
Distribución de jugadores por clusters (gráfico de la izquierda)

Distribución de Posiciones de jugadores por Clusters (gráfico de la derecha)



Distribución de Clusters por Equipos

- Se muestra la distribución de Clusters dentro de cada equipo.
- Es interesante ya que, dejando a un lado la clusterización realizada, esperarías que lo "normal" fuera que cada equipo tuviera una distribución adecuada de jugadores segun sus características.
- Matemáticamente, el algoritmo de K-Means a pesar de no agrupar de forma "balanceada" a los jugadores en su Cluster, logra percibir como por cada equipo existe por lo menos un jugador en alguno de estos Clusters de características similares



Bibliografías

- https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.shapiro.html
- https://pingouin-stats.org/build/html/generated/pingouin.multivariate_normality.html
- https://365datascience.com/tutorials/python-tutorials/pca-k-means/
- https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html
- https://365datascience.com/tutorials/python-tutorials/k-means-clustering/