

Taller 1

Simón García A00371828

1. ¿Qué tipo de variable es cada una de las columnas? (recuerde responder según las 4 clasificaciones vistas en clase).

- Player_id = Nominal
- Player_name = Nominal
- Age = Discreta
- Country = Nominal
- Height_cm = Continua
- Weight_lbs = Continua
- Position = Nominal
- Overall = Discreta
- Potential = Discreta
- Value_millions = Continua
- Pref_foot = Nominal
- Weak_foot = Ordinal
- Skill_moves = Ordinal
- International_reputation = Ordinal
- Attacking_work_rate = Ordinal
- Defensive_work_rate = Ordinal
- Body_type = Nominal
- Club_country = Nominal
- Club_name = Nominal
- Club_overall = Discreta

2. Realice un completo análisis descriptivo de las columnas que vamos a utilizar:

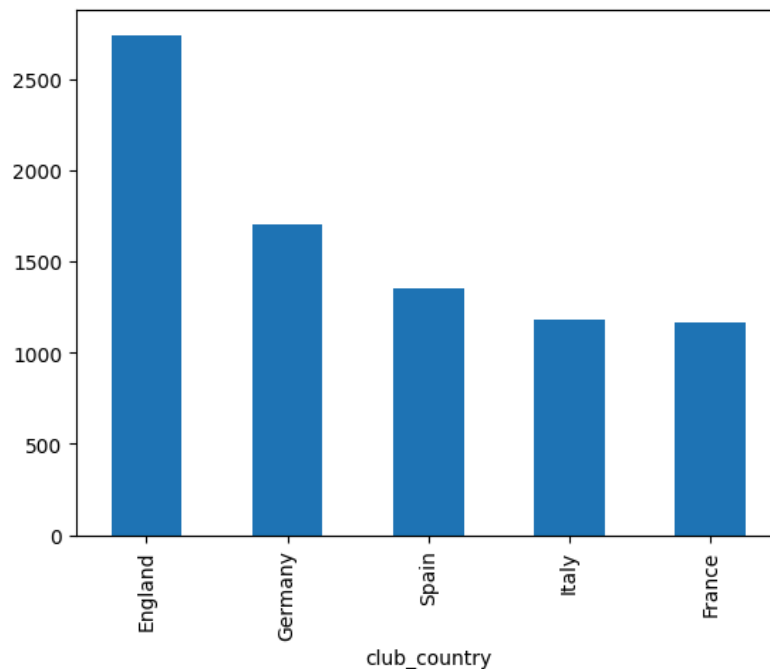
Se cuenta con 22080 datos ingresados

- **Age**
 - La edad va desde los 16 hasta los 44 años y un promedio de 24.89 años. El mínimo y máximo están bastante alejados del promedio mostrando a unos jugadores muy jóvenes y otros que están en la edad de retirarse. La desviación muestra que la variabilidad en la edad no es tan alta ya que con solo 4.89 años de edad significa que esos dos valores anteriores si están muy alejados de ese promedio.
- **height_cm**
 - La altura va desde los 155 cm hasta los 206 cm y un promedio de 181.39 cm. Esto revela que la mayoría de los futbolistas registrados son bastante altos (comparado con el promedio colombiano) y la variación en altura es alrededor de 7 cm.
- **position**

- Por mucho el rol más común es defensa central con 3876 representantes seguido por 3113 delanteros centrales. El rol menos elegido es el de mediapunta solo con 140 jugadores. La posición más común es la de defensores más arqueros con 9,799 seguido por 8024 mediocampistas y por último solo 4257 delanteros.

-
- **Club_country**

- Los jugadores vienen de clubes de 54 países. Los top 3 países representados son Inglaterra (2741), Alemania, y España. Los 3 menos representados son Islandia, Moroco y Ghana (5)



-
- **Overall**
 - El puntaje global de desempeño se encuentra en un 65% y muestra una desviación del 6.14%. Por un lado me parece que o los estándares están muy altos o el desempeño general no fue muy bueno durante el 2023. Si hay un máximo del 91% lo que significa que no es imposible tener un buen puntaje, pero el resto sí está bastante por debajo.
- **club_overall**
 - Comparando el desempeño de los jugadores con las de su equipo, si parece haber una consistencia ya que los equipos mostraron un desempeño del 68% con una desviación del 5%. Es decir que los resultados tampoco parecieron ser buenos a nivel general para los equipos tampoco. En este caso si se ve que el desempeño de ese jugador que sacó 91% es alto ya que incluso el desempeño más alto de un equipo fue solo 85%.
- **value_millions**
 - En cuanto el valor en el mercado de los jugadores hay una variabilidad relativamente baja en los datos, excepto que hay un gran número de valores

O



- O

Q

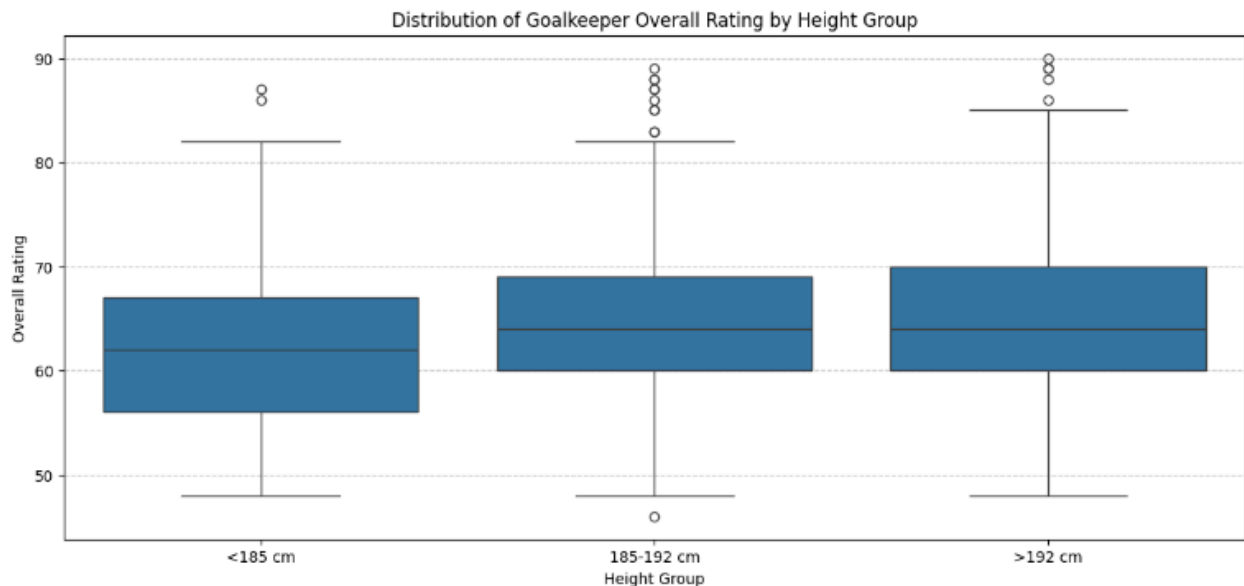
- Q



3. Un miembro del cuerpo técnico asegura que los arqueros (GK) altos tienen un mejor desempeño (overall). ¿Qué puede afirmar usted a través de un análisis descriptivo? (Nota: tome por definición tres grupos de estatura: i) menor a 185 cm; ii) entre 185 y 192 cm; iii) más de 192 cm)

	count	mean	std	min	25%	50%	75%	max
height_group								
<185 cm	649.0	62.092450	7.253400	48.0	56.0	62.0	67.0	87.0
185-192 cm	1264.0	64.481013	7.319944	46.0	60.0	64.0	69.0	89.0
>192 cm	517.0	64.746615	7.639378	48.0	60.0	64.0	70.0	90.0

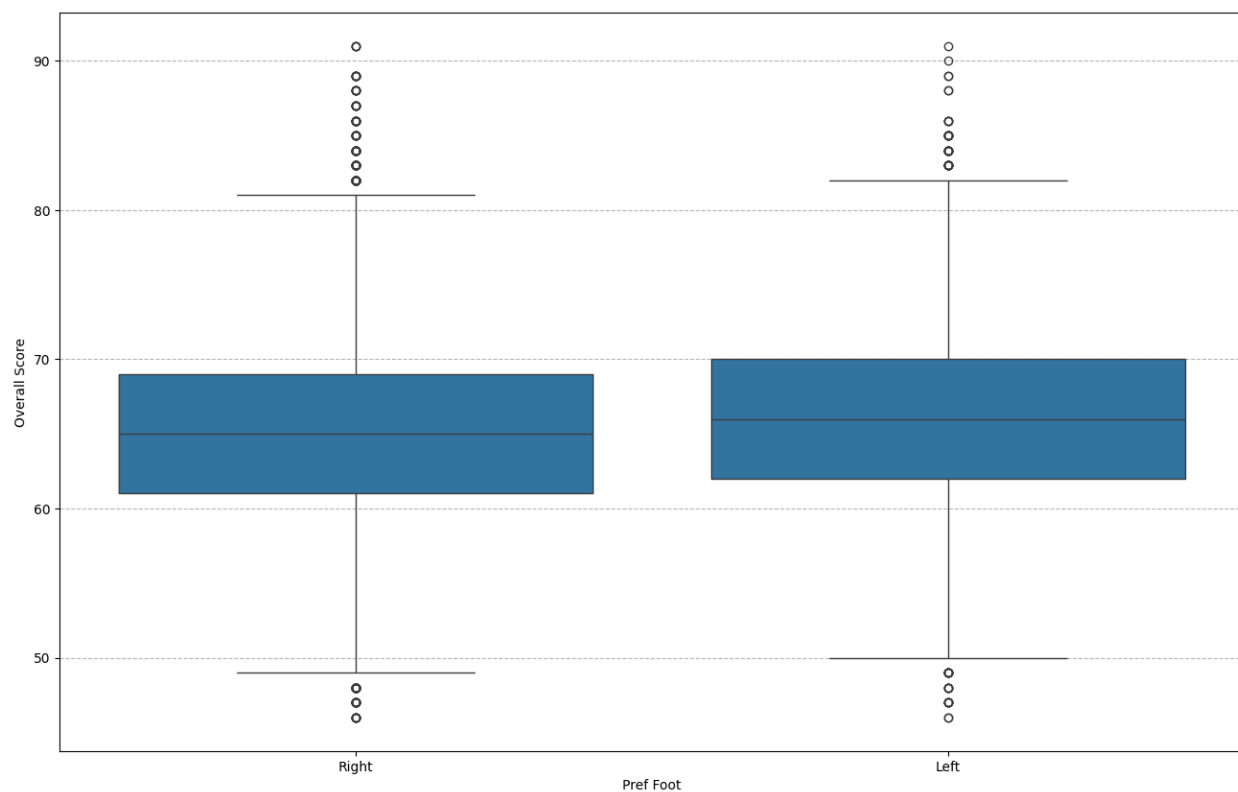
Con el análisis descriptivo se puede afirmar que los promedios están muy cercanos y teniendo en cuenta que la desviación estándar es de 7% probablemente significa que no hay una diferencia estadística entre los 3 puntajes overall de los 3 grupos tomados en consideración. Adicionalmente, el mínimo está entre 46 y 38, y el máximo entre 87 y 90. En general los resultados están muy cercanos, lo que se puede comprobar con la gráfica de barras siguiente. Es importante notar que incluso los cuartiles están sobrelapados, así que seguramente si se hace un ANOVA para verificar si la altura tiene un efecto en el desempeño los resultados probablemente serían desfavorables para la hipótesis del miembro del cuerpo técnico.



4. Un analista afirma que los jugadores zurdos, tienen un mejor desempeño (overall) que los diestros. ¿Qué puede decir sobre esta afirmación a través de un análisis descriptivo?

	count	mean	std	min	25%	50%	75%	max
pref_foot								
Left	5347.0	65.766037	6.548680	46.0	62.0	66.0	70.0	91.0
Right	16733.0	65.050738	6.875354	46.0	61.0	65.0	69.0	91.0

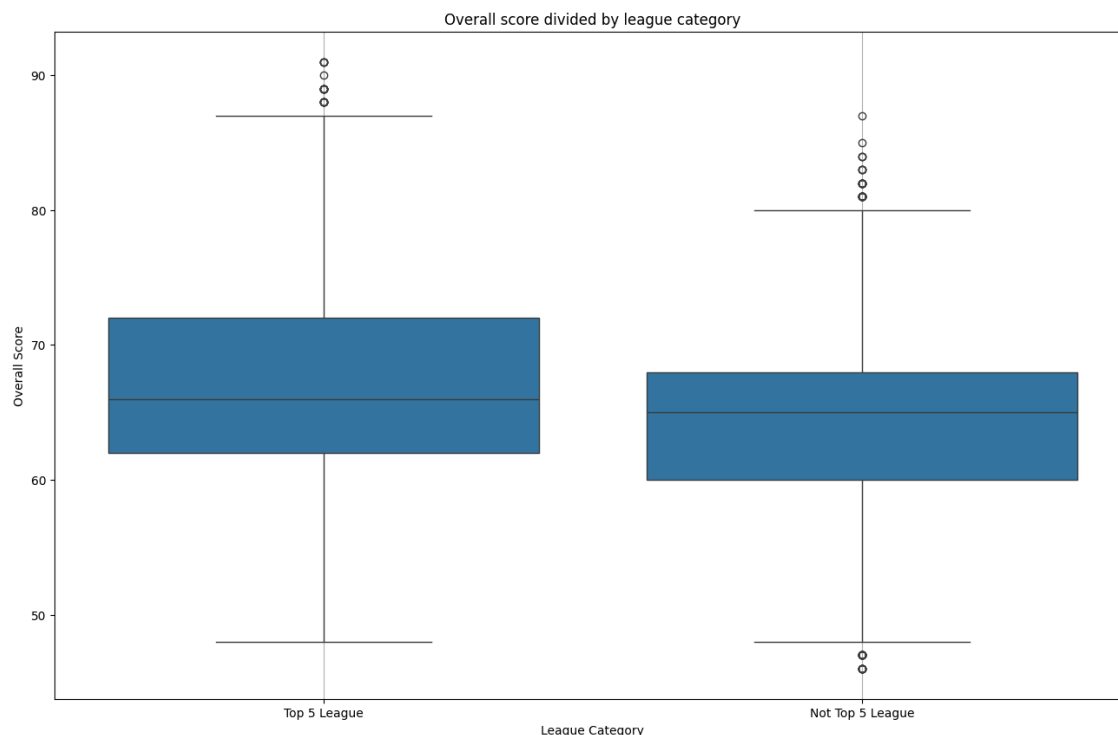
Viendo la distribución del puntaje general basado en la pierna dominante del jugador, primero que nada se puede observar un gran desbalance en el número de jugadores que prefiere su pierna izquierda de los que prefieren su pierna derecha. En ciertos casos ese desbalance puede llegar a afectar las pruebas estadísticas, pero teniendo en cuenta que son tantos datos probablemente no sea significativo. Por otro lado, el promedio entre los dos grupos, y la desviación estándar están muy cerca entre ellos. Al igual que el caso anterior, si se hace un ANOVA para definir si la pierna de preferencia tiene un efecto en el desempeño del jugador probablemente revelará que no existe tal diferenciación. No solo el mínimo y el máximo son iguales, si no que si se observa la figura posterior, el número y distribución de los datos atípicos se ve muy similar. Observando la figura también se puede notar que los bigotes de las cajas entre los dos grupos se sobrelapan lo que apoyaría que un análisis estadístico adicional saldría en contra de la hipótesis del analista.



5. Un scout asegura que los mejores jugadores, según el puntaje de desempeño (overall), juegan en clubes que pertenecen al grupo de “ligas top5” (Inglaterra, España, Alemania, Italia y Francia). ¿Qué puede decir sobre estas dos afirmaciones a través de un análisis descriptivo?

overall								
	count	mean	std	min	25%	50%	75%	max
leagueCategory								
Not Top 5 League	13939.0	64.277710	6.329658	46.0	60.0	65.0	68.0	87.0
Top 5 League	8141.0	66.844122	7.268095	48.0	62.0	66.0	72.0	91.0

Luego de agrupar a los jugadores dentro de la categoría de si están o no en un equipo de las ligas top5, se puede observar en el análisis descriptivo que aunque sean muchos más los equipos que se encuentran por fuera del top 5, la media sigue siendo muy cercana con una desviación estándar relativamente alta. Como la media y la desviación se están solapando, significa que probablemente un análisis estadístico diría que las dos categorías no tienen un efecto significativo en el desempeño de un jugador. Si se puede notar que hay una mínima diferencia comparando los mínimos y los máximos, pero no cambia el hecho que cuando se mira la distribución de los datos, solo hay un mínimo desplazamiento de los cuartiles y el bigote superior en la liga top 5 comparado a los que no hacen parte de esa categoría. Puede que haya un ínfima diferencia, pero probablemente no estadísticamente significativa.



6. Otro scout asegura que la proporción de jugadores zurdos es mayor en la posición de centro campista (CM) que en la de centro campista ofensivo (CAM). ¿Qué puede decir sobre esta afirmación?

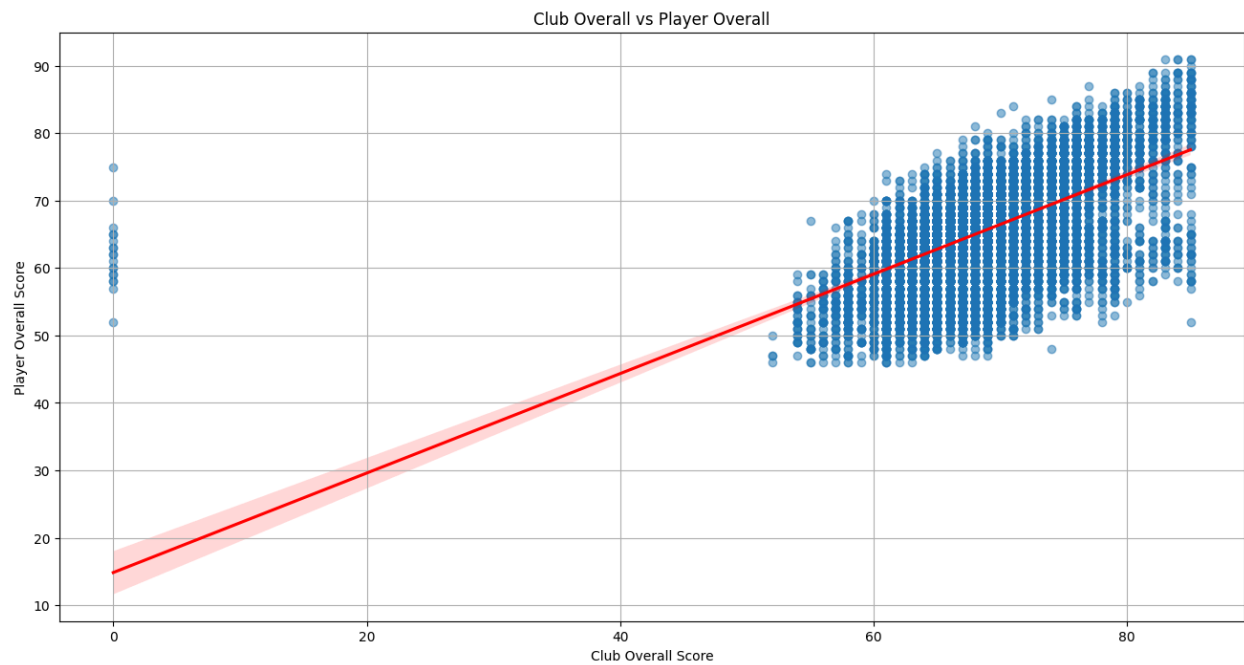
position	
LB	1281
CB	918
CM	481
ST	459
LM	356
CAM	353
RM	324
LWB	319
GK	272
CDM	224
RW	194
LW	118
CF	30
RB	12
RWB	6

Del total de 5347 que prefieren la pierna izquierda, 481 son centrocampistas mientras que 353 son centrocampistas ofensivos. Esta afirmación es cierta teniendo en cuenta que del grupo de zurdos los centrocampistas representan el 9.00% mientras que los centrocampistas ofensivos solo el 6.60% de los jugadores. En términos de todos los jugadores, los centrocampistas zurdos representan el 2.18% mientras que los centrocampistas ofensivos zurdos son el 1.60% de los jugadores en la base de datos.

7. Otro scout asegura que los mejores jugadores juegan en los mejores clubes del mundo ¿Es cierta esta afirmación de que existe una correlación entre el desempeño de los jugadores (overall)

	overall	club_overall
overall	1.000000	0.581029
club_overall	0.581029	1.000000

Como lo muestra la correlación, si hay una relación positiva entre el puntaje del club y el desempeño del jugador. Esta correlación no es tan fuerte, teniendo en cuenta que es menor al 60%, pero en un ámbito tan variable como lo es el fútbol sí tendría sentido aprobar este tipo de medidas con mayor variabilidad. Si se ve la gráfica inferior, se puede notar que la línea que representa la regresión tiene una tendencia positiva. Aunque al inicio los clubes que no tienen una puntuación aumentan la variabilidad al inicio, los jugadores en sus clubes siguen teniendo un desempeño similar al resto de jugadores lo que significa que probablemente el dato estaba faltando y fue llenado con 0's. Ya solo teniendo en cuenta los datos de jugadores con clubes con puntajes adecuados, se ve un tipo de dispersión que tiene una mayor prevalencia a la parte inferior a la recta de la regresión. Esto probablemente se debe a que los puntajes altos eran mucho menos comunes que los puntajes más bajos.



Info Adicional:

Nomenclatura posiciones en el campo		
Abreviatura	Rol	Posición
GK	Arquero	Arquero
RWB	Carrilero Derecho	Defensor
RB	Lateral Derecho	Defensor
CB	Defensa Central	Defensor
LB	Lateral Izquierdo	Defensor
LWB	Carrilero Izquierdo	Defensor
CDM	Medio Centro Defensivo	Mediocampista
RM	Medio Derecho	Mediocampista
CM	Medio Centro	Mediocampista
LM	Medio Izquierdo	Mediocampista
CAM	Medio Centro Ofensivo	Mediocampista
RF	Segundo Delantero Derecho	Delantero
CF	Media Punta	Delantero
LF	Segundo Delantero Izquierdo	Delantero
RW	Extremo Derecho	Delantero
ST	Delantero Centro	Delantero
LW	Extremo Izquierdo	Delantero
COLUMNA	DESCRIPCIÓN	
_player_id	Id del jugador	
_player_name	Nombre del jugador	
_age	Edad	
_country	Nacionalidad	
_height_cm	Estatura en cm	
_weight_lbs	Peso en lbs	
_position	Posición	
_overall	Puntaje global del desempeño	
_potential	Puntaje potencial del talento	
_Value_millions	Valor de mercado en millones de dólares	
_pref_foot	Pierna preferida	
_weak_foot	Habilidad con la pierna no preferida (1 a 5 siendo 5 lo más alto)	
_skill_moves	Habilidad para realizar maniobras con el balón (1 a 5 siendo 5 lo más alto)	
_international_reputation	Reputación internacionl (1 a 5, siendo 5 lo más alto)	
_Attacking_work_rate	Disposición a ir al ataque	
_Defensive_work_rate	Disposición a defender	
_club_country	País del club en el que juega	
_club_name	Nombre del club	
_club_overall	Puntaje global del desempeño del club	