Proyecto EDA

ANÁLISIS DE RENDIMIENTO EN JUGADORES DE LA RFEF TEMPORADA 23/24.

Nicolas Matías Muñiz Escalada – Bootcamp Data Science

Objetivos:

- Analizar rendimientos de jugadores pertenecientes a la liga RFEF, de la temporada 23-24.
- Ver el comportamiento de los jugadores en sus equipos según sus características principales, contestando distintas preguntas.

Procedimiento:

- El análisis estará segmentado en 3 partes:

1. Limpieza del Data Set.
2. Análisis Univariante y Bivariante.
3. Conclusiones y respuestas.

Información

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10107 entries, 0 to 10106
Data columns (total 12 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   id                10107 non-null  int64
 1   nombre            10107 non-null  object
 2   partidos_jugados  10107 non-null  int64
 3   partidos_titular  10107 non-null  int64
 4   goles             10107 non-null  int64
 5   asistencias       10107 non-null  int64
 6   tarjeta_amarilla  10107 non-null  int64
 7   edad              10107 non-null  int64
 8   altura            10107 non-null  float64
 9   valor             10107 non-null  float64
 10  rating            10107 non-null  int64
 11  posicion          10107 non-null  object
dtypes: float64(2), int64(8), object(2)
memory usage: 947.7+ KB
```

- DATA SET LIBRE DE VALORES NULOS
- ESTA COMPUESTO POR 12 COLUMNAS Y 10107 FILAS.
- LA MAYORIA SON COLUMNAS NUMERICAS

Preguntas a resolver:

- 1. Se relaciona la estatura del jugador por la posición en la que juega? Y la edad?
- 2. Hay una tendencia de mala conducta por posición?
- 3. Los delanteros mejor calificados son los que tienen mas goles? Si hacen mas goles, tienen mas partidos de titular?

Preguntas a resolver:

- 4. Se opta generalmente por tener jugadores experimentados dentro del campo?

- 5. Mostrar los mejores 5 jugadores por posición y ver en que grupo de edad se encuentran.

COMENZAMOS…

Vista del Data Set

	id	nombre	partidos_jugados	partidos_titular	goles	asistencias	tarjeta_amarilla	edad	altura	valor	rating	posicion
0	0	Alberto Varo	34	34	21	1	2	31	191.000000	174640.0	59	Portero
1	2	Dani Parra	5	4	3	0	1	24	188.000000	91220.0	44	Portero
2	4	Joan Oriol	37	36	1	3	9	37	175.000000	119000.0	62	Lateral Izquierdo
3	5	P. Trigueros	37	37	5	1	6	31	188.000000	217500.0	57	Defensa Central
4	6	Nacho González	31	31	2	0	8	29	185.000000	340340.0	61	Defensa Central
...	...	...	...	...	...	...	...	...	...	...	...	...
10102	12529	A. Lopez	13	4	0	0	1	0	181.820179	1280.0	40	Mediocentro
10103	12530	Xesc Navalón	32	27	19	0	3	26	171.000000	24590.0	48	Delantero Centro
10104	12532	Juanca	19	12	6	0	3	30	180.344828	43710.0	47	Extremo Izquierdo
10105	12533	Álex Sánchez	30	22	11	0	2	19	180.883984	10760.0	42	Delantero Centro
10106	12534	Roberto	21	4	1	0	3	22	180.000000	777.8	26	Mediocentro

10107 rows × 12 columns

PRESENTACION DE VARIABLES

Columna/Variable	Descripción	Tipo	Cat/Num
Nombre	Nombre del jugador	String	Numerica Continua
partidos_jugados	Cantidad de partidos jugados en la temporada 2023 - 2024	Int	Numerica Discreta
partidos_titular	Cantidad de partidos que el jugador entro de titular	Int	Numerica Discreta
goles	Cantidad de goles que ha marcado en la temporada	Int	Numerica Discreta
asistencias	Cantidad de asistencias que dio el jugador	Int	Numerica Discreta
tarjeta_amarilla	cantidad de tarjetas amarillas recibidas	Int	Numerica Discreta
edad	Edad del jugador	Int	Numerica Discreta
altura	Altura del jugador	Float	Numerica Continua
valor	En Euros, cuanto vale el jugador	Float	Numerica Continua
rating	Puntaje medio del jugador en la temporada	Int	Numerica Discreta
posicion	Posicion del jugador	String	Categorica

LIMPIEZA

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10107 entries, 0 to 10106
Data columns (total 11 columns):
 #   Column            Non-Null Count  Dtype  
---  ------            --------------  -----  
 0   nombre            10107 non-null  object 
 1   partidos_jugados  10107 non-null  int64  
 2   partidos_titular  10107 non-null  int64  
 3   goles             10107 non-null  int64  
 4   asistencias       10107 non-null  int64  
 5   tarjeta_amarilla  10107 non-null  int64  
 6   edad              10107 non-null  int64  
 7   altura            10107 non-null  float64
 8   valor             10107 non-null  float64
 9   rating            10107 non-null  int64  
 10  posicion          10107 non-null  object 
dtypes: float64(2), int64(7), object(2)
memory usage: 868.7+ KB
```

- ELIMINAMOS LA COLUMNA ID YA QUE AL PARECER ES UN INDICE QUE TIENE VALORES DESORDENADOS Y NO NOS SIRVE PARA EL ANALISIS QUE QUEREMOS HACER.

LIMPIEZA

```
<class 'pandas.core.frame.DataFrame'>
Index: 8765 entries, 0 to 10106
Data columns (total 11 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   nombre             8765 non-null   object
 1   partidos_jugados   8765 non-null   int64
 2   partidos_titular   8765 non-null   int64
 3   goles              8765 non-null   int64
 4   asistencias        8765 non-null   int64
 5   tarjeta_amarilla   8765 non-null   int64
 6   edad               8765 non-null   int64
 7   altura             8765 non-null   float64
 8   valor              8765 non-null   float64
 9   rating             8765 non-null   int64
 10  posicion           8765 non-null   object
dtypes: float64(2), int64(7), object(2)
memory usage: 821.7+ KB
```

- LAS FILAS CON EDAD 0 REPRESENTAN UN 13% DEL DATAFRAME, Y COMO QUEREMOS RESPONDER UNA DE LAS PREGUNTAS CON CERTEZA NECESITAMOS LAS EDADES REALES.

LIMPIEZA

posicion	
Mediocentro	1729
Defensa Central	1623
Delantero Centro	1387
Portero	1027
Lateral Derecho	619
Lateral Izquierdo	595
Extremo Derecho	495
Extremo Izquierdo	461
MP	413
Mediocentro Defensivo	199
MI	99
MD	93
CAI	12
CAD	6
PT	4
MPI	2
MPD	1

Name: count, dtype: int64

- LA COLUMNA POSICION TIENE VALORES QUE QUIEREN DECIR LO MISMO PERO CON LAS INICIALES.

posicion	
Mediocampista	1921
Defensor central	1822
Delantero Centro	1387
Delantero extremo	1372
Defensor lateral	1232
Portero	1031

Name: count, dtype: int64

- LO SIMPLIFICO Y LO AGRUPO SEGÚN LAS POSICIONES BASICAS. A LO QUE QUIERO RESPONDER ES MAS SIMPLE VERLO DE ESTA MANERA.

LIMPIEZA

- POR ULTIMO, CONVIENE CREAR DOS COLUMNAS PARA CONTESTAR LAS PREGUNTAS QUE QUEREMOS RESOLVER:

```
grupo_edad
Jovenes     3537
Novatos     3195
Adultos     1671
Mayores      362
Name: count, dtype: int64
```

CATEGORIZAR LAS EDADES DE LOS JUGADORES

```
grupo_altura
Medios     7000
Bajos      1105
Altos       660
Name: count, dtype: int64
```

CATEGORIZAR LAS ALTURAS DE LOS JUGADORES

NUEVO DATA SET

```
<class 'pandas.core.frame.DataFrame'>
Index: 8765 entries, 0 to 10106
Data columns (total 13 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   nombre            8765 non-null   object
 1   partidos_jugados  8765 non-null   int64
 2   partidos_titular  8765 non-null   int64
 3   goles             8765 non-null   int64
 4   asistencias       8765 non-null   int64
 5   tarjeta_amarilla  8765 non-null   int64
 6   edad              8765 non-null   int64
 7   altura            8765 non-null   float64
 8   valor             8765 non-null   float64
 9   rating            8765 non-null   int64
 10  posicion          8765 non-null   object
 11  grupo_edad        8765 non-null   object
 12  grupo_altura      8765 non-null   object
dtypes: float64(2), int64(7), object(4)
memory usage: 958.7+ KB
```

- Ahora el Data Frame tiene 13 columnas nuevamente pero con 8765 filas porque quitamos las edades = 0.
- Ya puedo realizar el análisis variable a variable.

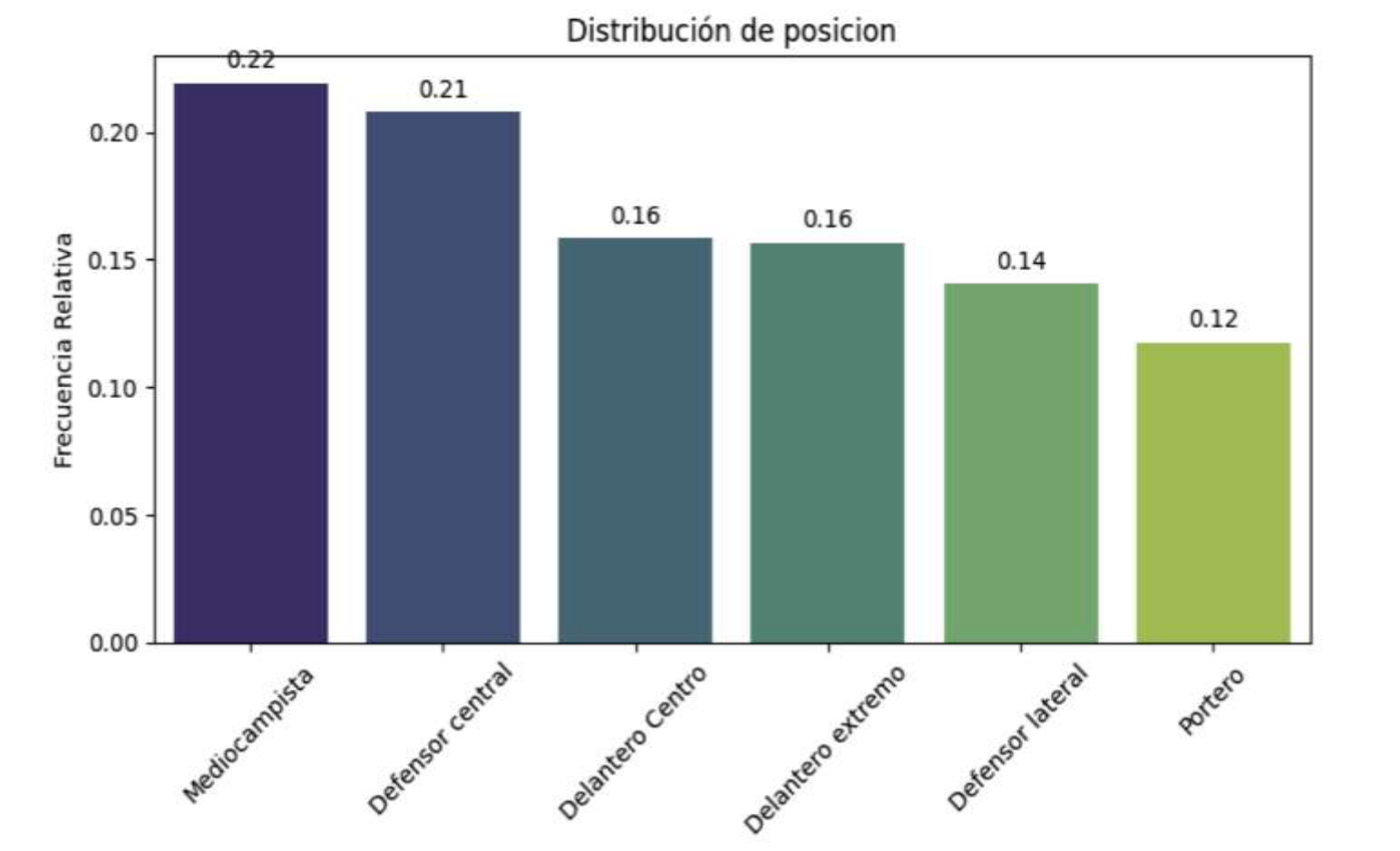
ANALISIS UNIVARIANTE

- Cuales son varibales numéricas o categóricas según su cardinalidad.

	Card	%_Card	Tipo	tipo_sugerido
nombre	7436	84.837422	object	Numerica continua
partidos_jugados	36	0.410724	int64	Numerica discreta
partidos_titular	39	0.444952	int64	Numerica discreta
goles	61	0.69595	int64	Numerica discreta
asistencias	13	0.148317	int64	Numerica discreta
tarjeta_amarilla	19	0.216771	int64	Numerica discreta
edad	31	0.353679	int64	Numerica discreta
altura	3357	38.300057	float64	Numerica continua
valor	4598	52.458642	float64	Numerica continua
rating	45	0.513406	int64	Numerica discreta
posicion	6	0.068454	object	Categorica
grupo_edad	4	0.045636	object	Categorica
grupo_altura	3	0.034227	object	Categorica

Categoricas

- Posición:

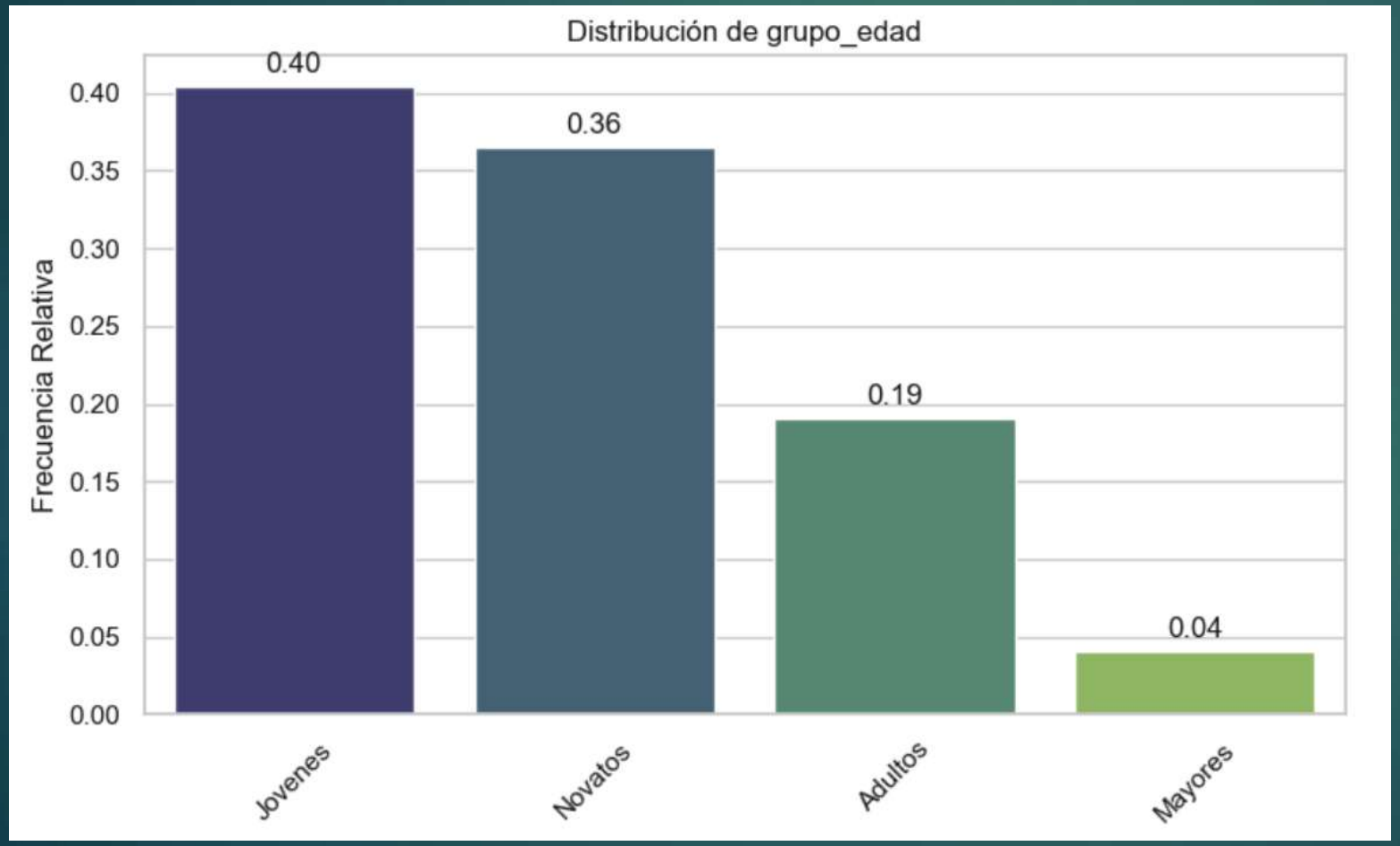


Se muestran las frecuencias para las posiciones.

Se esperaba que para los porteros sea el que menor proporción haya porque hay solo una posición por cubrir.

Categóricas

- Grupo_edad:

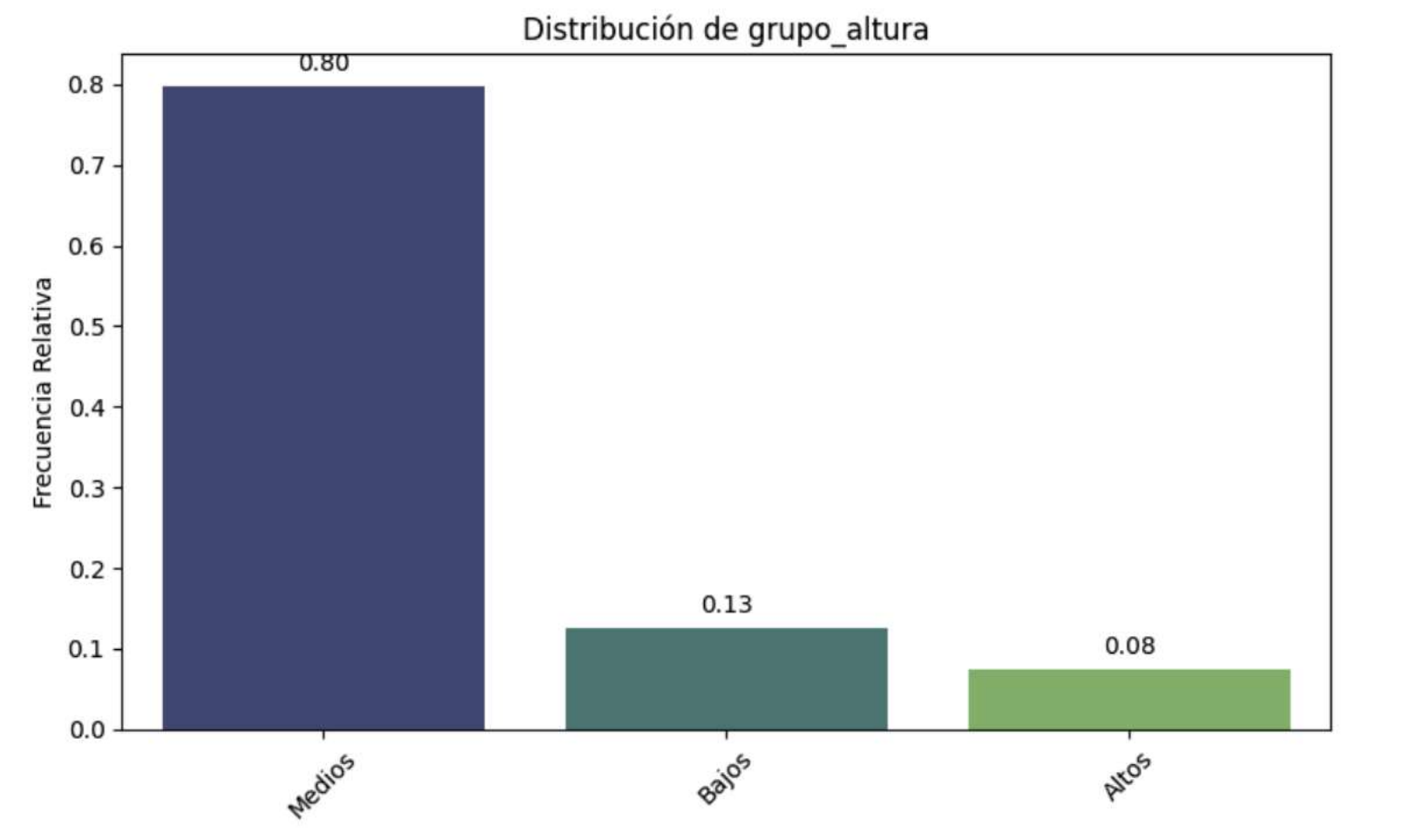


- Novatos: 16 – 22 años
- Jóvenes: 23 – 28 años
- Adultos: 29 – 35 años
- Mayores: +36 años

Se esperaba esta distribución en la frecuencia. Y aquí es donde se plantea la hipótesis de: Para que un mayor continue su carrera tiene que jugar la mayoría de partidos, caso contrario dejaría el deporte profesionalmente.

Categóricas

- Grupo_altura:



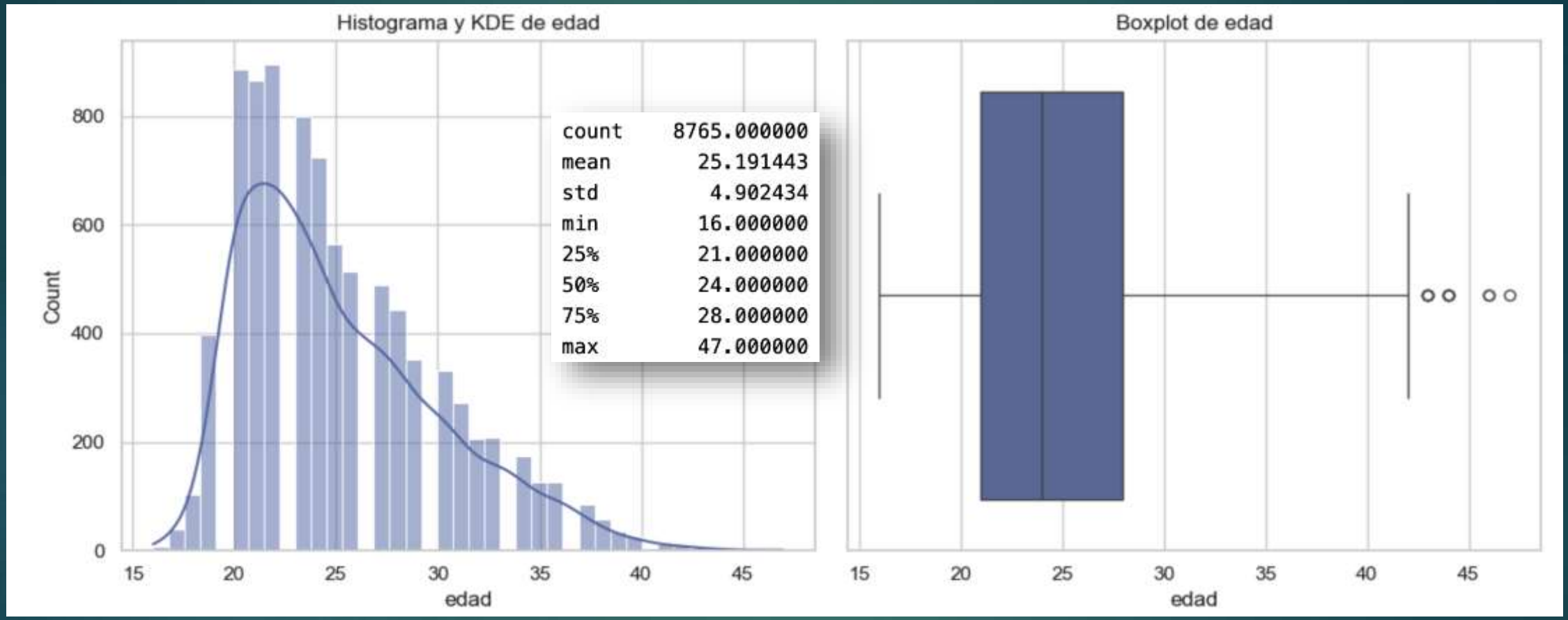
- Bajos: 160 – 175 cm
- Medios: 175 – 187 cms
- Altos: +187 cms

Se esperaba que las alturas de los jugadores sean normales, por lo tanto de estatura media.

Esta variable nos servirá para contestar si las alturas influyen en las posiciones, pero sabemos que el 80% son medios.

Numericas

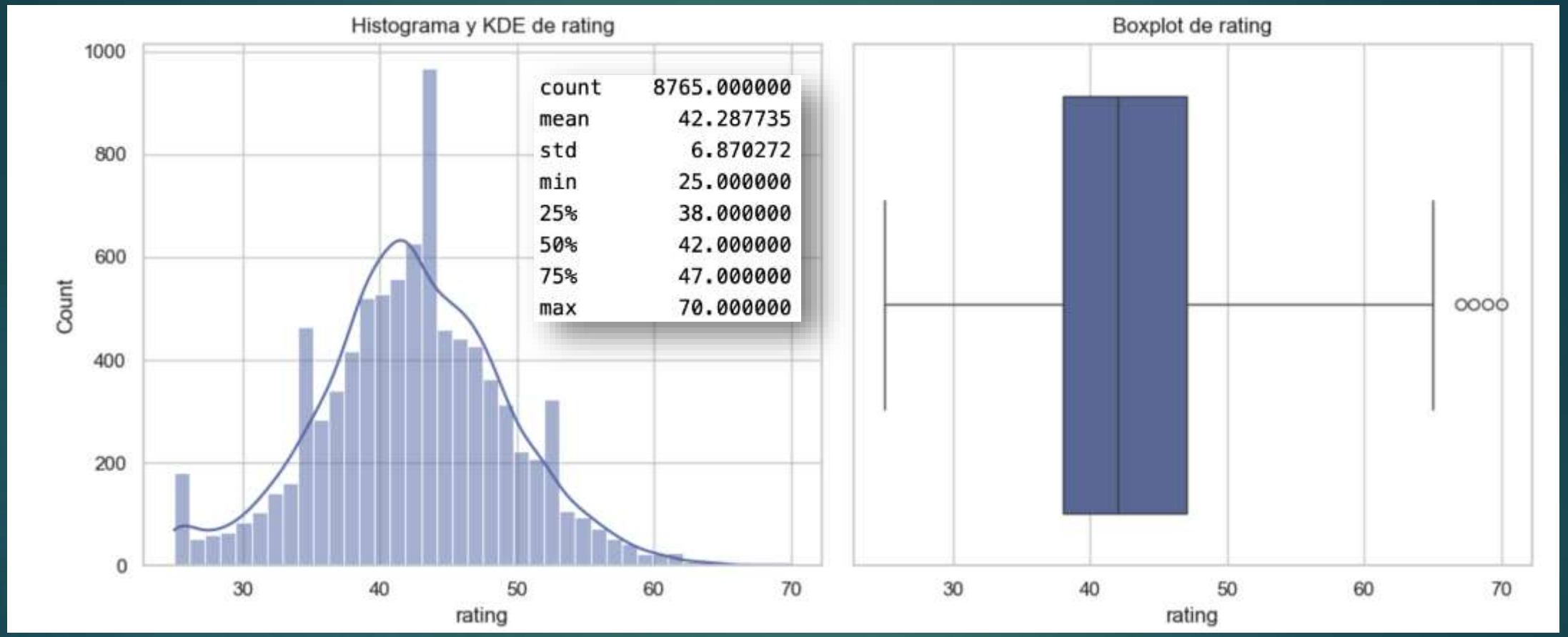
- edad:



La distribución del grafico es bastante normal, con un rango de min 16 y max 47 años. Pero la mayor concentración de los valores se da entre 21 y 28 años de edad. Ademas existen jugadores que ya pasaron su edad y van hasta los 47 identificando a partir de los 41 que ya es un Outlier.

Numericas

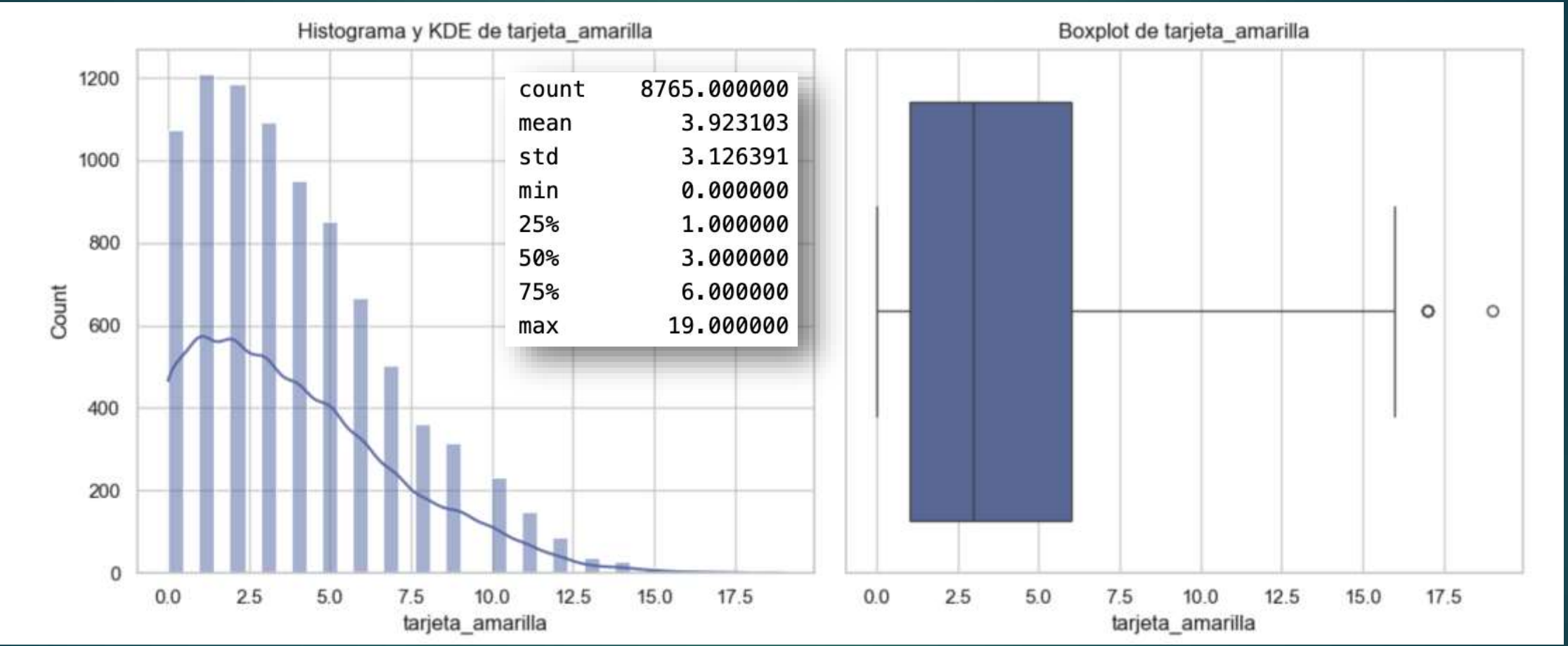
- rating:



Su distribución también es normal, y la mayor cantidad de valores se centran entre 38 y 47 puntos. La mediana y la media son prácticamente iguales. Hay existencia de Outliers que los podemos tratar como "ejemplos" a la hora de ver que se tiene en cuenta en el puntaje.

Numericas

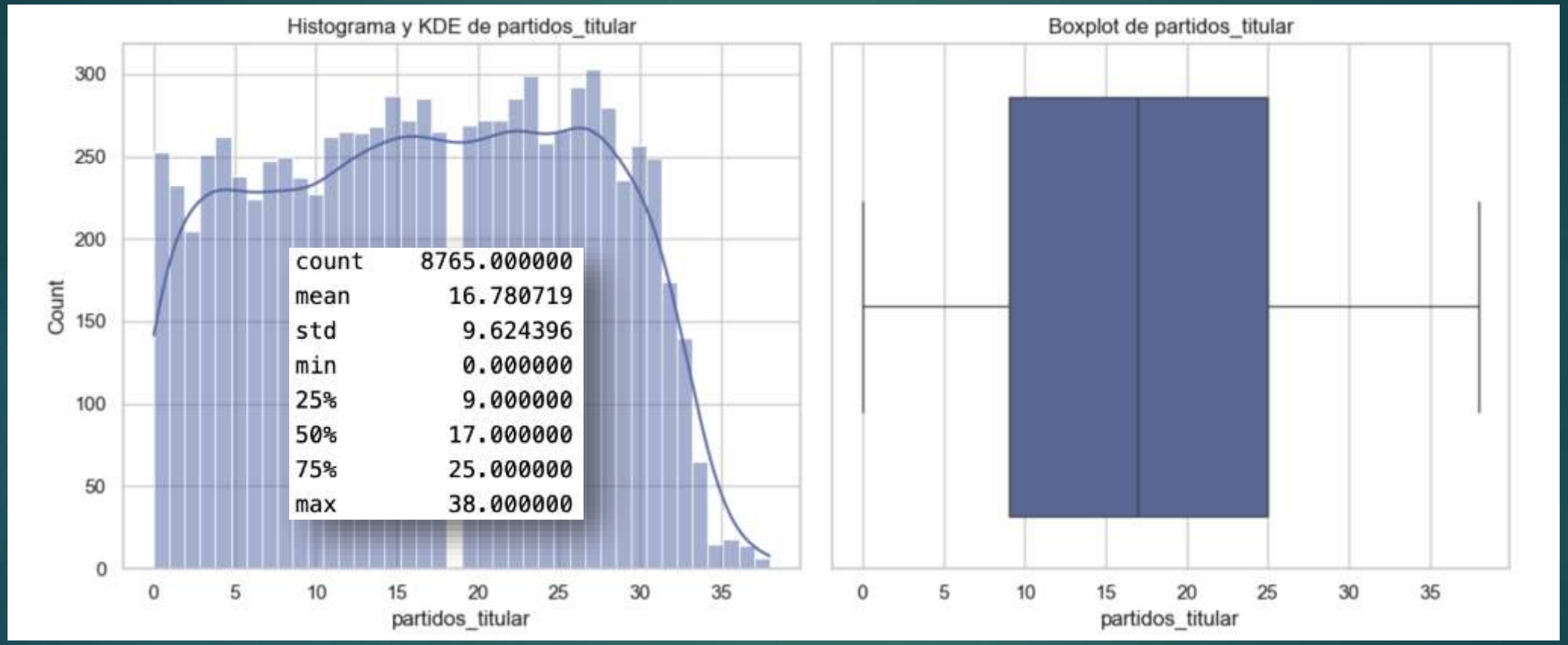
- Tarjetas_amarillas:



La conducta de los jugadores presenta una media y mediana que difieren solo en un punto, la distribución de los valores se concentra entre 1 y 6 tarjetas amarillas por jugador. Y hay existencia de outliers, por lo podrian plantearse advertencias a los clubes para mantener la conducta en la liga.

Numericas

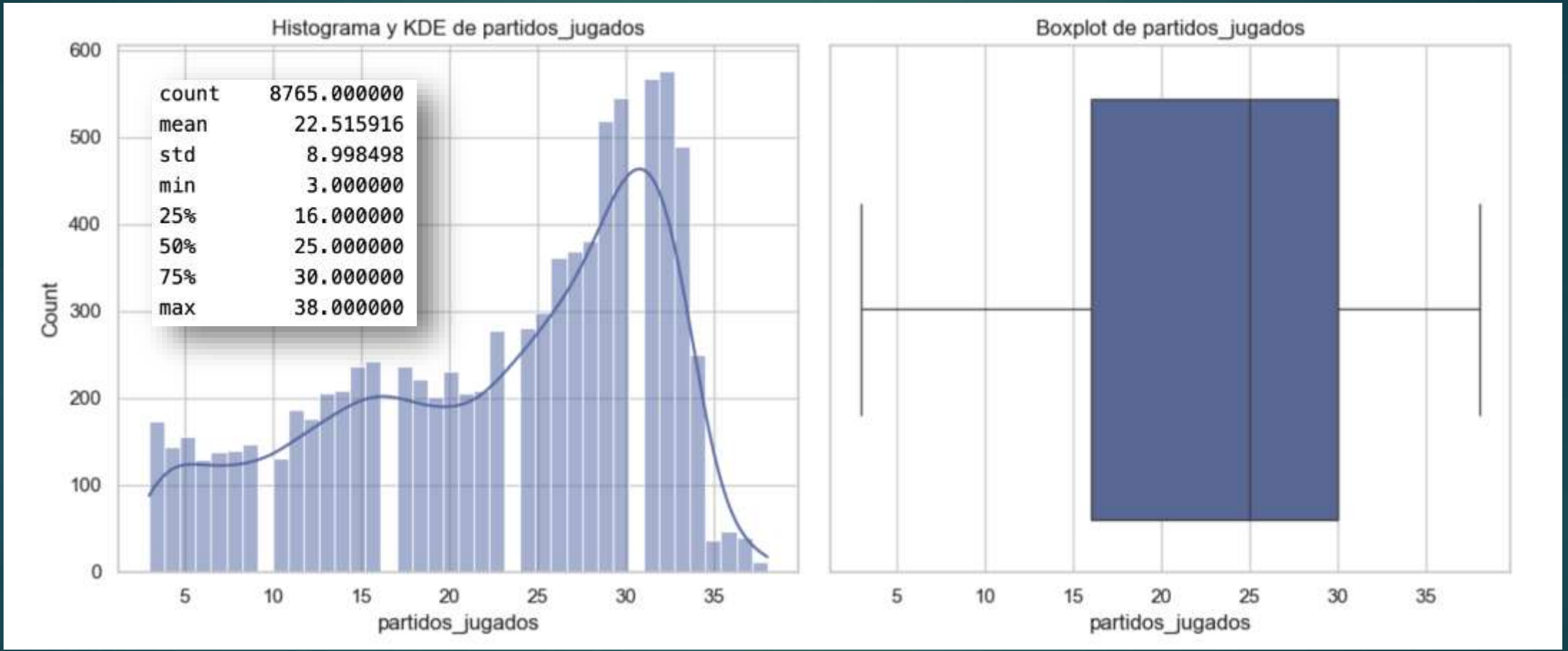
- Partidos_titular:



Esta variable nos puede orientar en que podrían ser la consecuencia de un buen rating o rendimiento, ya que a mayor cantidad de partidos de titular se puede decir que el jugador es muy bueno y fundamental para su equipo, lo contrario nos daría un redimiendo mas bajo. Se mostrará en el análisis bivariante.

Numericas

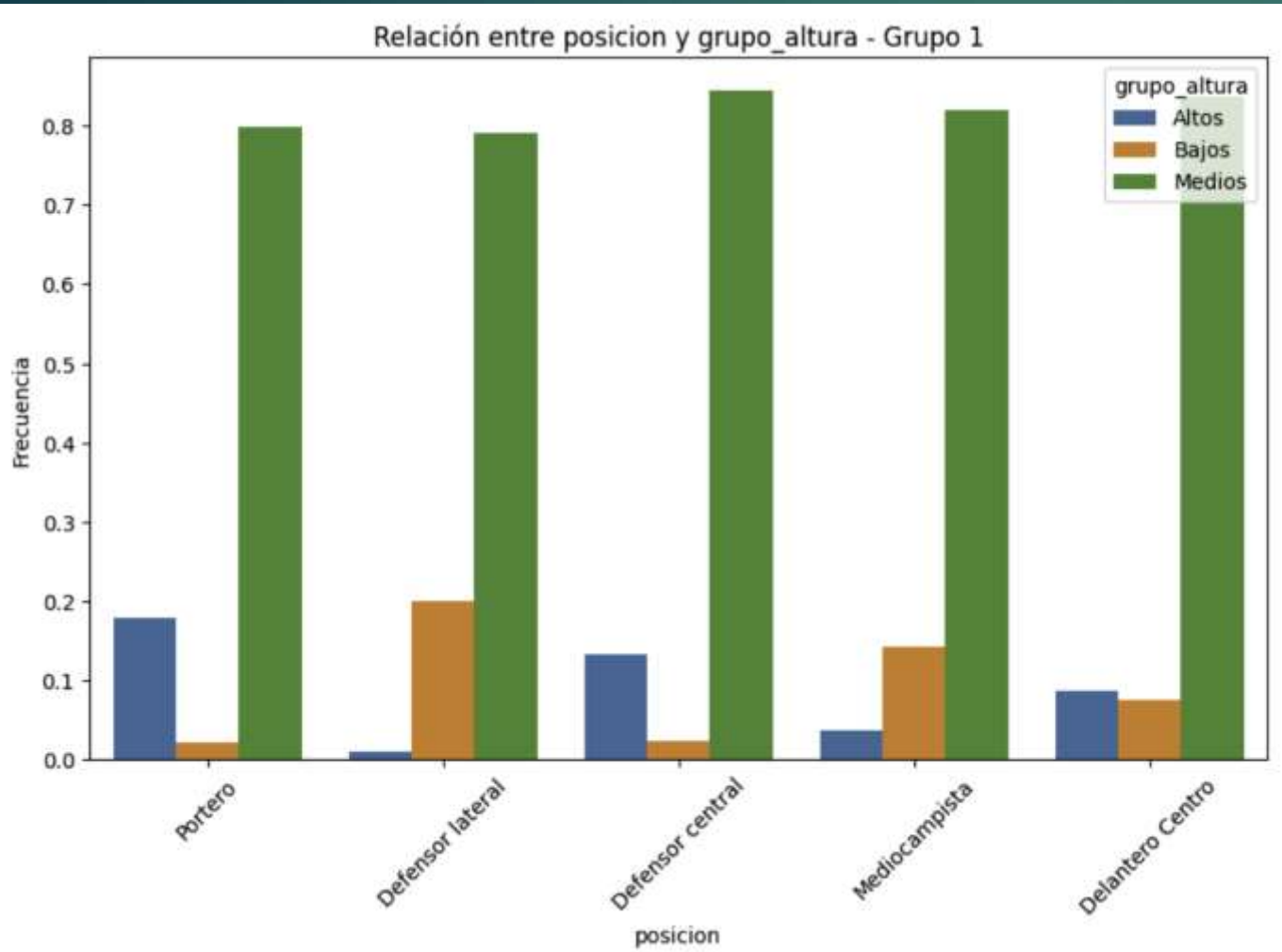
- partidos_jugados:

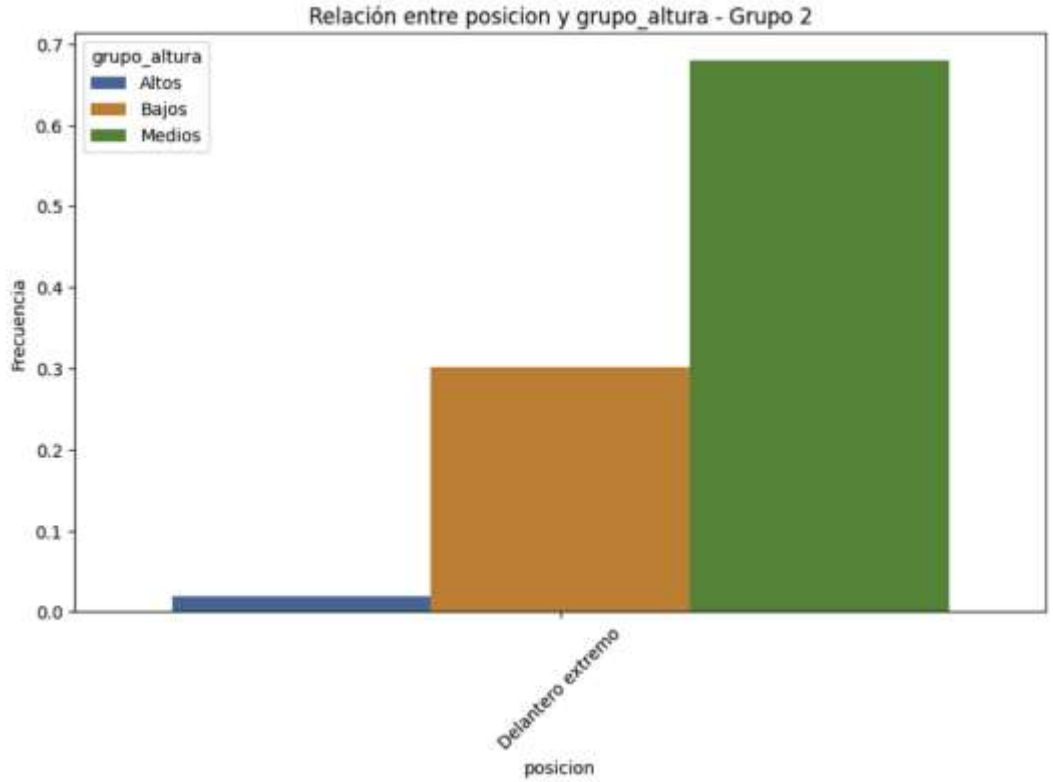


Analisis bivariante

- A partir de ahora ya puedo ir contestando las preguntas que al principio quería resolver. Por lo tanto recordare las preguntas y las respondere luego de la combinacion de las variables.

- 1. Se relaciona la estatura del jugador por la posición en la que juega, teniendo en cuenta que los defensores, delanteros y porteros pueden ser mas altos que los mediocampistas? Y que pasa con la edad?





RESPUESTA 1 – (Altura)

Efectivamente, recordando que la estatura media es la que mas predomina en el data set, quiero observar solo los de estatura alta y baja. Y se reparten de la siguiente manera:

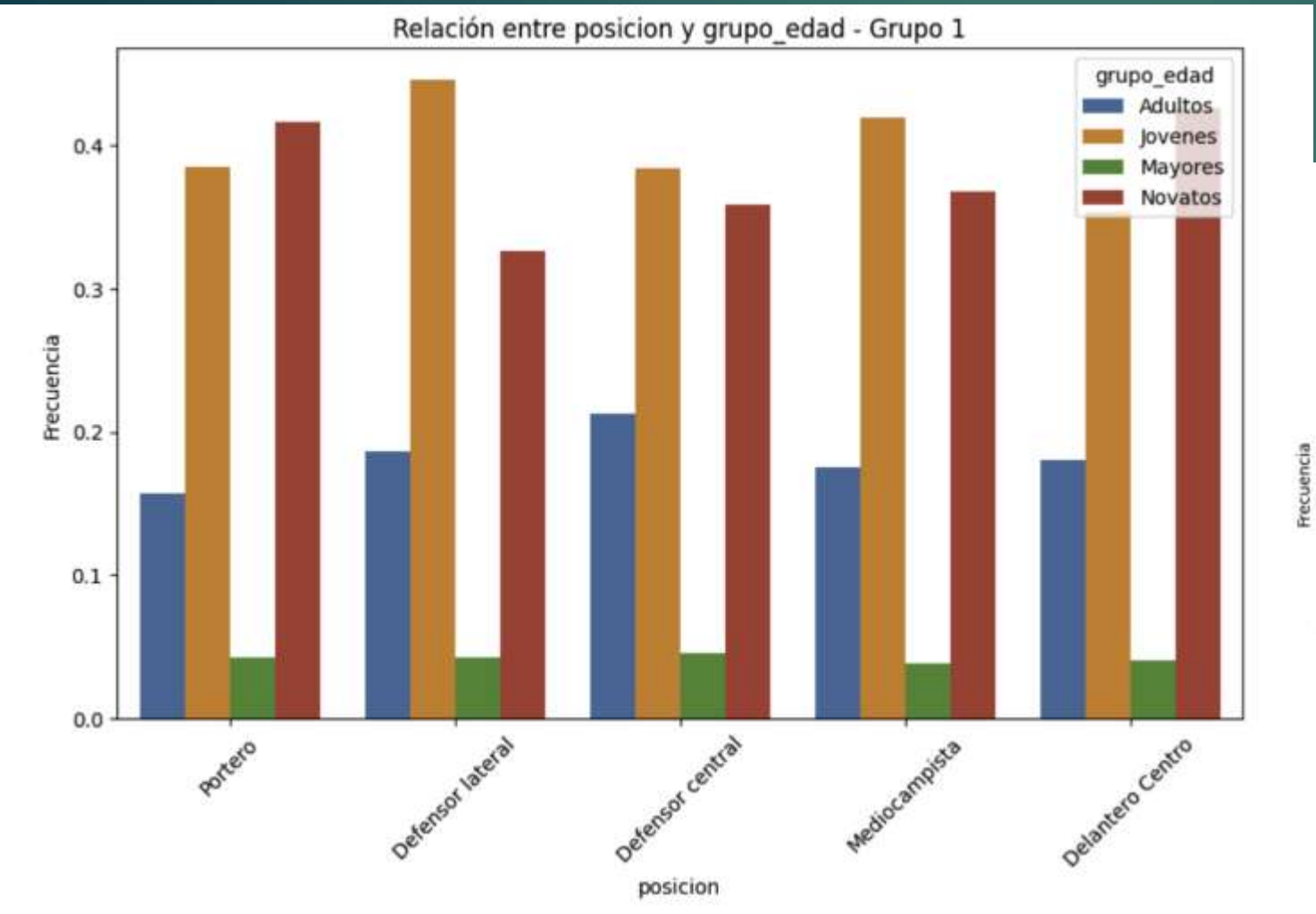
- Al parecer es excluyente si no eres alto para jugar de portero, defensor central, o delantero centro.

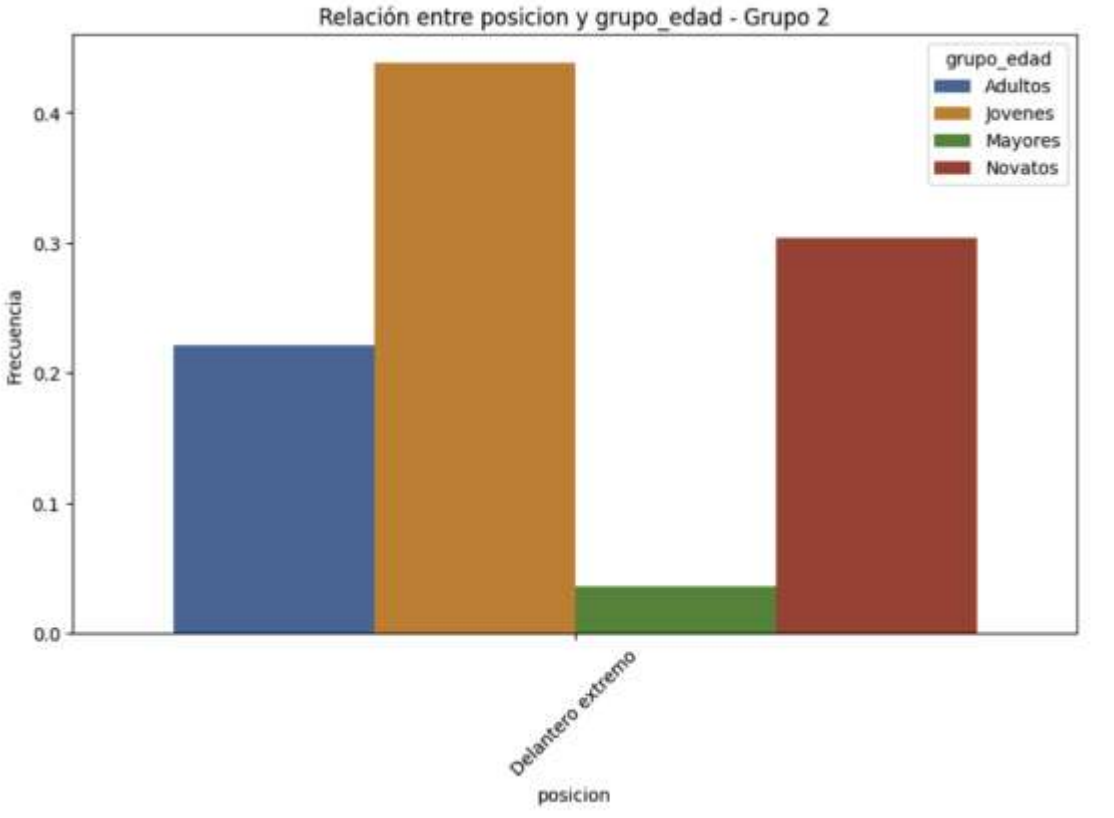
- Por otro lado, si eres bajo es recomendable ser delantero extremo, defensor lateral o mediocampista.

Siguiendo este análisis si eres jugador y te interesa probarte en un club, deberías tener en cuenta en que posición probarte según tu altura, a no ser que seas extraordinario en este deporte y puedas jugar en cualquier posición sin importar cuanto midas.

1. Y que pasa con la edad?

Según los gráficos, vemos que la edad no parece definir en que posición conviene jugar, pero si vemos que tal vez si eres mayor no optes por jugar de delantero extremo, que es donde se necesita velocidad y explosión.



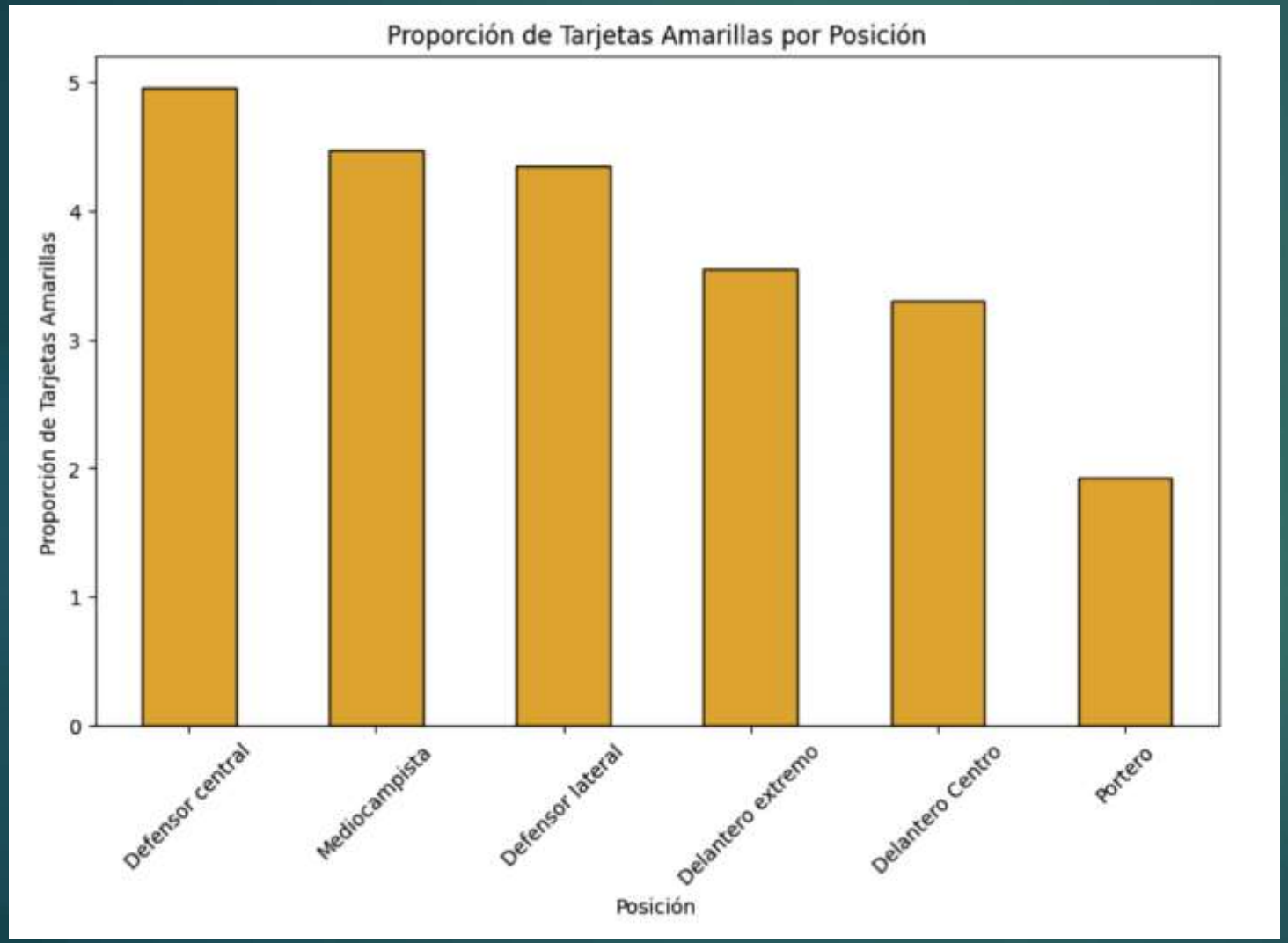


- 2. Hay una tendencia mas de mala conducta por posición?

Para responder esta pregunta utilizo ANOVA ya que estoy analizando una variable categórica no binaria con una numérica.

- Valor p: 8.669552376579864e-173
- Esto significa que rechazamos la hipótesis nula y decimos que hay posiciones que reciben mas tarjetas amarillas que otras.
- Lo veo con el grafico de la media respecto a cada una.

- 2. Hay una tendencia de mala conducta por posición?

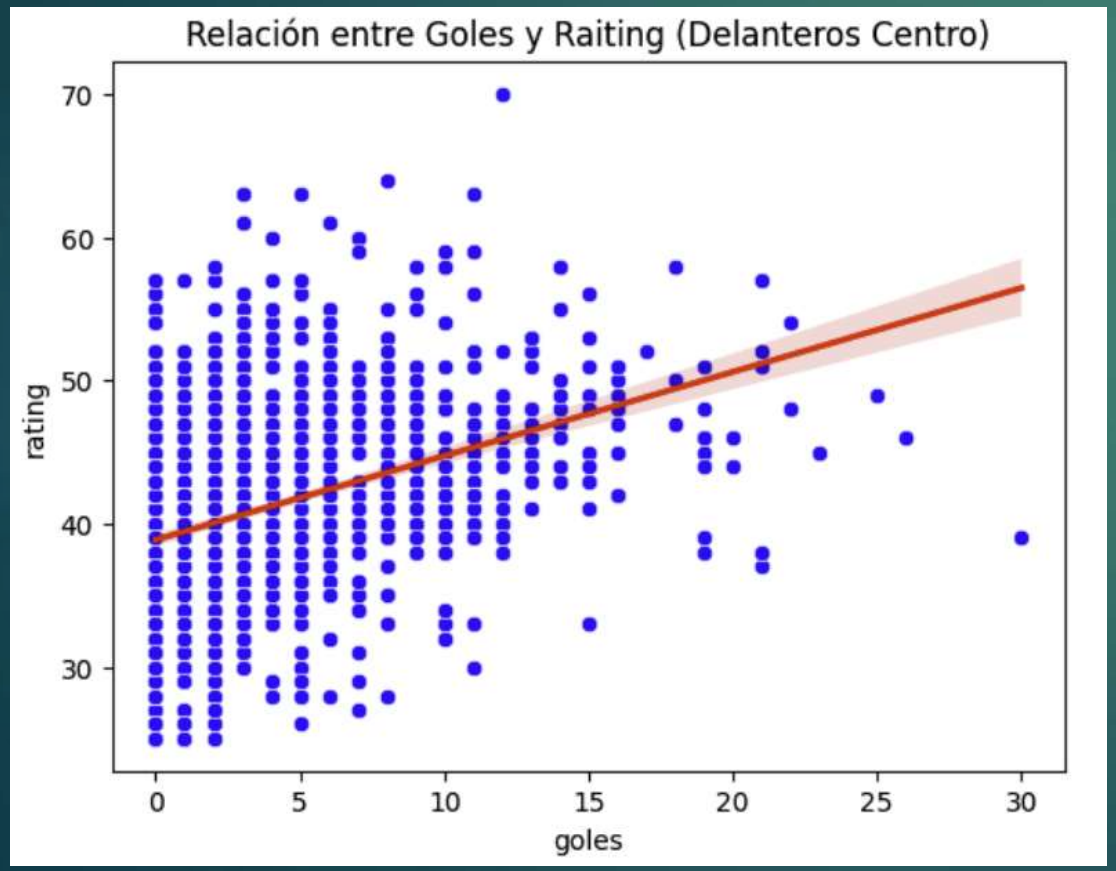


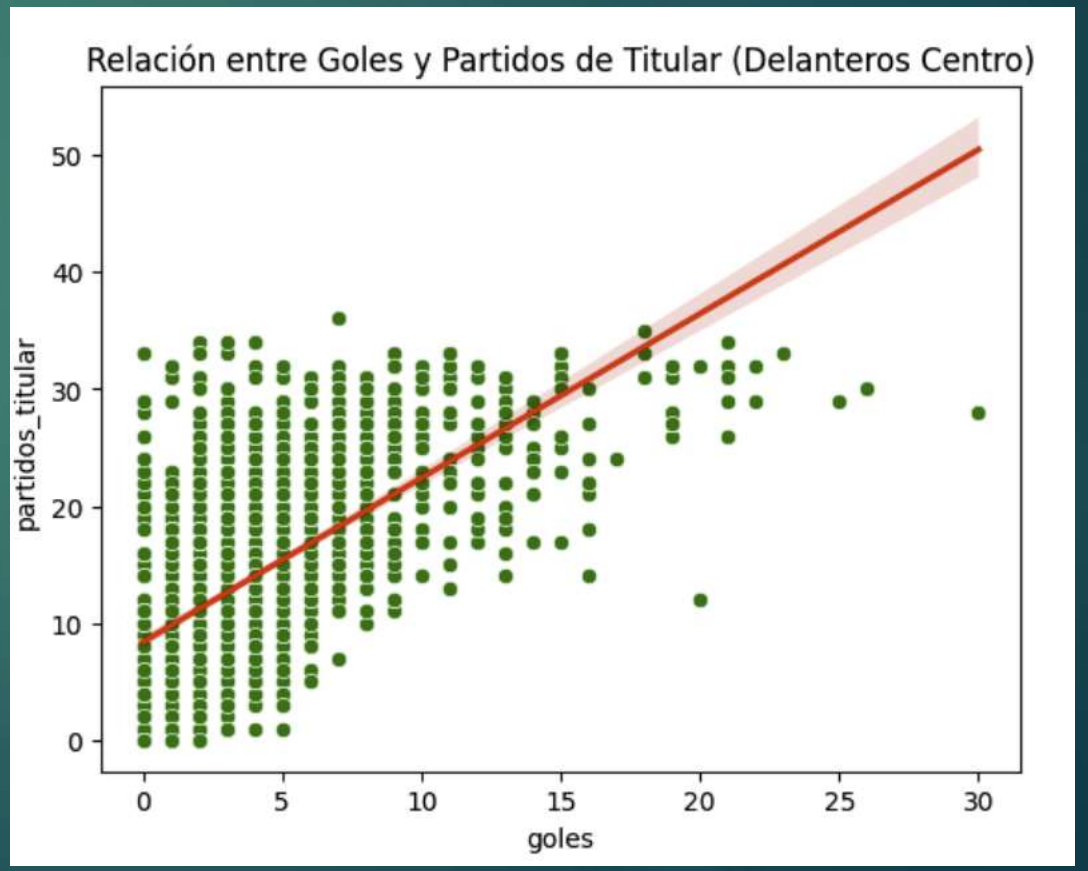
Es lógico pensar que los que juegan detrás de la mitad del campo son los que mas interrumpen el juego, por lo tanto los defensores centrales, mediocampistas y defensores laterales son los que mas amonestaciones reciben.

Se esperaba en los porteros que la media sea baja, ya que ellos no cometen muchas infracciones, se propone para otros análisis ver cuantas amarillas fueron por mala conducta y cuantas fueron por interrumpir una jugada con falta.

- 3. Los delanteros centro mejor calificados son los que tienen mas goles? Si hacen mas goles, tienen mas partidos de titular?

El grafico nos muestra una gran dispersión en ambos casos, porque no es excluyente que haga un gol o que juegue de titular. Pero es interesante ver como si haces mas goles, mas partidos de titular juegas. Esto era lo esperado ya que la función principal de un delantero centro hoy en día es hacer goles, aunque, si analizamos rating, la correlación es un poco mas leve. Esto se da a que hay otros factores a la hora de analizar el rating del jugador.



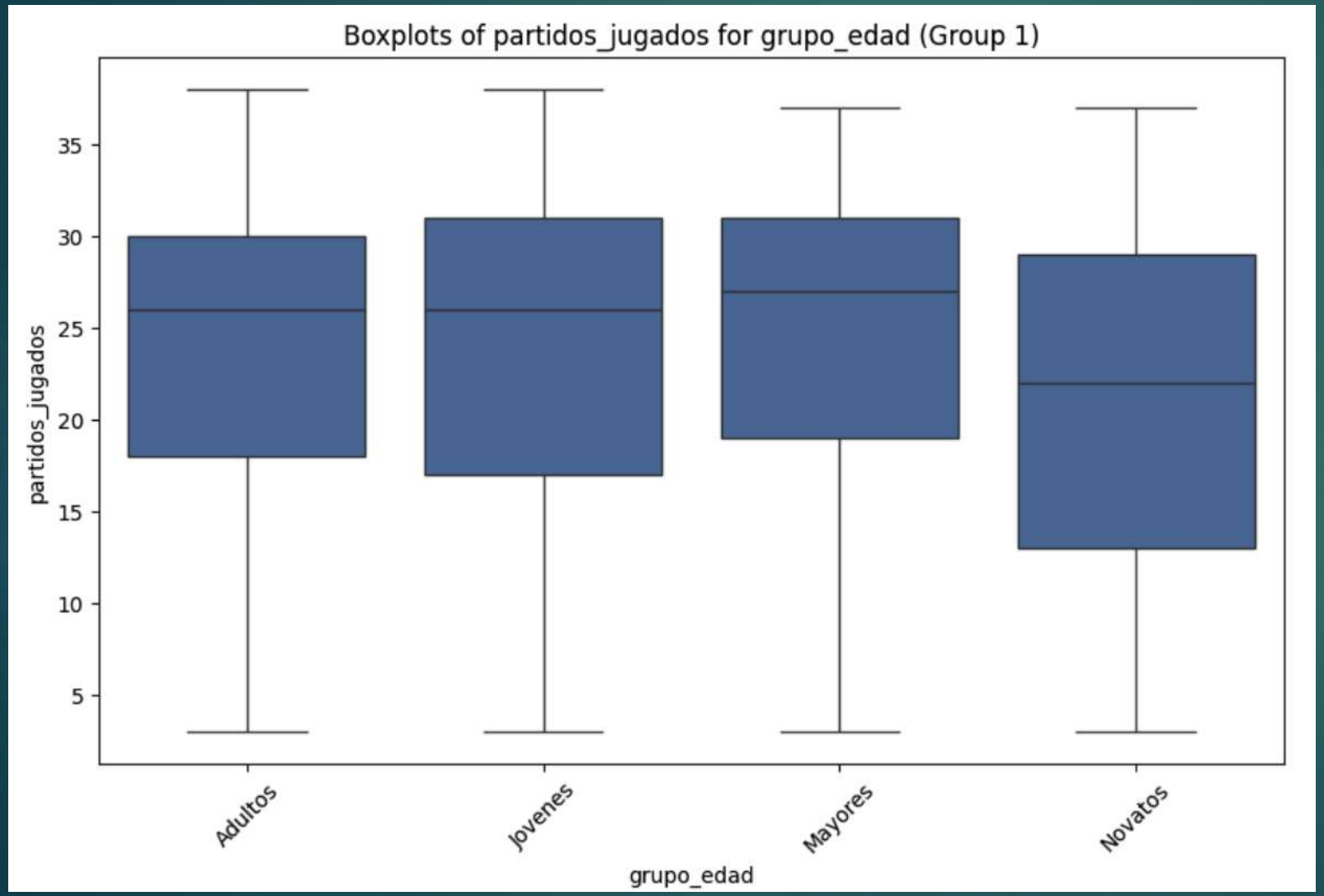


- 4. Se opta generalmente por tener jugadores experimentados dentro del campo?

Nuevamente utilizo ANOVA (grupo_edad = categorica, partidos_jugados = numérica)

- Valor p: 5.029069877438021e-52
- Esto significa que rechazamos la hipótesis nula y las medias en partidos jugados son distintas.
- Ahora lo intentare explicar con un grafico de boxplot.

- 4. Se opta generalmente por tener jugadores experimentados dentro del campo?



Es interesante ver este grafico ya que hemos visto antes que este Data Set esta minado de Novatos y Jovenes.

No obstante, los mayores se ubican por encima de la media y no importa que sean el grupo de menor frecuencia, por lo que esto indica que los mayores que quedan en la liga son los que mas partidos juegan de los distintos grupos de edad.

- 5. Mostrar los mejores 5 jugadores por posición y ver en que grupo de edad se encuentran.

	nombre	partidos_jugados	partidos_titular	goles	asistencias	tarjeta_amarilla	edad	altura	valor	rating	posicion	grupo_edad
325	Einar Galilea	24	22	2	0	5	30	185.0	1310000.0	69	Defensor	Adultos
305	J. Martínez	8	7	0	0	1	31	191.0	2140000.0	68	Defensor	Adultos
353	A. Escassi	31	30	4	1	12	35	186.0	368700.0	65	Defensor	Adultos
769	David Andújar	16	16	1	0	5	32	189.0	619420.0	64	Defensor	Adultos
792	Pablo Vázquez	36	36	4	0	7	29	189.0	818290.0	64	Defensor	Adultos
805	Lucas Pérez	31	31	12	17	6	35	180.0	991770.0	70	Delantero	Adultos
343	Dioni	36	24	8	1	3	34	184.0	424600.0	64	Delantero	Adultos
434	Pedro León	30	20	3	8	5	37	183.0	1090000.0	64	Delantero	Mayores
642	Emilio Nsue	30	24	11	3	4	34	181.0	395480.0	63	Delantero	Adultos
782	Borja Valle	19	17	3	3	6	31	177.0	1040000.0	63	Delantero	Adultos
802	Salva Sevilla	18	14	1	1	1	40	178.0	115660.0	67	Mediocampista	Mayores
427	Tomás Pina	28	20	1	0	10	36	185.0	497370.0	64	Mediocampista	Mayores
545	Juanan	37	27	2	5	6	34	183.0	217510.0	62	Mediocampista	Adultos
311	Yussi Diarra	37	35	5	4	6	25	174.0	533770.0	61	Mediocampista	Jovenes
383	Cristian Rodríguez	16	15	3	7	3	28	175.0	574840.0	61	Mediocampista	Jovenes
324	A. Herrero	38	38	25	0	3	30	183.0	561970.0	65	Portero	Adultos
351	B. Reynet	9	9	8	0	0	33	185.0	362390.0	62	Portero	Adultos
417	Manu García	30	30	24	0	5	33	192.0	215560.0	61	Portero	Adultos
1014	Miguel Bañuz	31	31	21	0	3	31	188.0	250390.0	60	Portero	Adultos
0	Alberto Varo	34	34	21	1	2	31	191.0	174640.0	59	Portero	Adultos

- 5. Mostrar los mejores 5 jugadores por posición y ver en que grupo de edad se encuentran.

	partidos_jugados	partidos_titular	goles	asistencias	tarjeta_amarilla	edad	altura	valor	rating
count	20.000000	20.000000	20.000000	20.000000	20.000000	20.000000	20.000000	2.000000e+01	20.000000
mean	26.950000	24.000000	7.950000	2.550000	4.650000	32.450000	183.950000	6.351180e+05	63.800000
std	9.428261	9.130631	8.293783	4.198684	2.924938	3.394655	5.276313	4.876168e+05	2.948684
min	8.000000	7.000000	0.000000	0.000000	0.000000	25.000000	174.000000	1.156600e+05	59.000000
25%	18.750000	16.750000	2.000000	0.000000	3.000000	30.750000	180.750000	3.343900e+05	61.750000
50%	30.000000	24.000000	4.000000	1.000000	5.000000	32.500000	184.500000	5.155700e+05	64.000000
75%	34.500000	31.000000	11.250000	3.250000	6.000000	34.250000	188.250000	8.616600e+05	65.000000
max	38.000000	38.000000	25.000000	17.000000	12.000000	40.000000	192.000000	2.140000e+06	70.000000

Según el rating, vemos los jugadores con mayor puntuación por posición y hay un dato muy curioso a tener en cuenta.

Los mejores jugadores por posición son los que tienen mas experiencia, por lo tanto si soy un equipo de estas competiciones, tendría que pensar en tenerlos dentro del campo de juego.

Muchas gracias !