1. Fuente y Procedencia del Dataset

La fuente de datos para el siguiente trabajo pertenece a un detallado contenido de atributos para cada jugador de futbol registrado en la edición del juego para consolas FIFA 19. el cual se encuentra en:

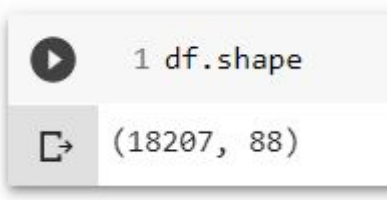
https://www.kaggle.com/karangadiya/fifa19

2. Motivos para Analizar Dataset

Desde que Alemania logró consagrarse como campeón del torneo Mundial FIFA en Brasil 2014 con el uso de BigData, siempre ha sido importante analizar las estadísticas de los jugadores de fútbol para poder armar un buen equipo para poder enfrentar a sus rivales en los diferentes torneos de futbol. Siempre se analiza el desempeño de los equipos, sus jugadores, sus métricas de goles y pases para poder plantear con qué alineamiento voy a enfrentar a mi siguiente rival.

3. Dimensiones del Dataset

Las dimensiones del dataset son 18207 filas y 88 columnas



4. Medidas Estadísticas Básicas del Dataset

El dataset contiene 42 columnas numéricas de importancia, para el trabajo vamos a tomar solamente 11 que son las siguientes:

Age (Edad), HeadingAccuracy (Eficacia con la cabeza), Dribbling(Habilidad para Eludir al oponente), BallControl (Control del balon), Acceleration(Aceleración), Agility(Agilidad), SprintSpeed (Velocidad de Sprint), Vision, Penalties(Conversion de Penales), Composure(Eficacia del jugador), Marking (Marcación)

Los valores de las columnas por analizar contiene valores entre 1 y 100 excepto la edad.

```
1 df['Age'].describe()
2
```

```
count    18207.000000
mean        25.122206
std          4.669943
min         16.000000
25%         21.000000
50%         25.000000
75%         28.000000
max         45.000000
Name: Age, dtype: float64
```

```
[143] 1 df['HeadingAccuracy'].describe()
```

```
count    18159.000000
mean        52.298144
std         17.379909
min          4.000000
25%         44.000000
50%         56.000000
75%         64.000000
max         94.000000
Name: HeadingAccuracy, dtype: float64
```

```
[ ]  1 df['Dribbling'].describe()
```

```
count    18159.000000
mean        55.371001
std         18.910371
min          4.000000
25%         49.000000
50%         61.000000
75%         68.000000
max         97.000000
Name: Dribbling, dtype: float64
```

```
[145]  1 df['BallControl'].describe()
       2
```

```
count    18159.000000
mean        58.369459
std         16.686595
min          5.000000
25%         54.000000
50%         63.000000
75%         69.000000
max         96.000000
Name: BallControl, dtype: float64
```

```
[146]  1 df['Acceleration'].describe()
       2
```

```
count    18159.000000
mean        64.614076
std         14.927780
min         12.000000
25%         57.000000
50%         67.000000
75%         75.000000
max         97.000000
Name: Acceleration, dtype: float64
```

```
[147]  1 df['Agility'].describe()
       2
```

```
count    18159.000000
mean        63.503607
std         14.766049
min         14.000000
25%         55.000000
50%         66.000000
75%         74.000000
max         96.000000
Name: Agility, dtype: float64
```

```
[148]  1 df['SprintSpeed'].describe()
```

```
count    18159.000000
mean        64.726967
std         14.649953
min         12.000000
25%         57.000000
50%         67.000000
75%         75.000000
max         96.000000
Name: SprintSpeed, dtype: float64
```

```
[149]  1 df['Vision'].describe()
       2
```

```
count    18159.000000
mean        53.400903
std         14.146881
min         10.000000
25%         44.000000
50%         55.000000
75%         64.000000
max         94.000000
Name: Vision, dtype: float64
```

```
[150]  1 df['Penalties'].describe()
       2
```

```
count    18159.000000
mean        48.548598
std         15.704053
min          5.000000
25%         39.000000
50%         49.000000
75%         60.000000
max         92.000000
Name: Penalties, dtype: float64
```

```
[151]  1 df['Composure'].describe()
       2
```

```
count    18159.000000
mean        58.648274
std         11.436133
min          3.000000
25%         51.000000
50%         60.000000
75%         67.000000
max         96.000000
Name: Composure, dtype: float64
```

```
[152]  1 df['Marking'].describe()
```

```
count    18159.000000
mean        47.281623
std         19.904397
min          3.000000
25%         30.000000
50%         53.000000
75%         64.000000
max         94.000000
Name: Marking, dtype: float64
```

5. Missing Values

Existen valores nulos

```
[170]  1 df.isna().sum().sort_values(ascending=False)
```

```
Club                241
Marking              48
Composure            48
Penalties            48
Vision               48
SprintSpeed          48
Agility              48
Acceleration         48
BallControl          48
Dribbling            48
HeadingAccuracy      48
Age                   0
Nationality           0
Name                  0
ID                    0
dtype: int64
```

Vemos que casi todos los valores de las habilidades que vamos a analizar menos la edad tienen valores nulos.

```
1 df['HeadingAccuracy'].sort_values(ascending=False)
```

```
204      94.0
102      94.0
700      93.0
818      93.0
499      93.0
         ...
13279     NaN
13280     NaN
13281     NaN
13282     NaN
13283     NaN
Name: HeadingAccuracy, Length: 18207, dtype: float64
```

Mostramos todos los nulos

```
[171] 1 df.loc[df['HeadingAccuracy'].isnull()]
```

	ID	Name	Nationality	Club	Age	HeadingAccuracy	Dribbling	BallControl	Acceleration	Agility	SprintSpeed	Vision	Penalties	Composure	Marking
13236	177971	J. McNulty	Scotland	Rochdale	33	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
13237	195380	J. Barrera	Nicaragua	Boyacá Chicó FC	29	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
13238	139317	J. Stead	England	Notts County	35	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
13239	240437	A. Semprini	Italy	Brescia	20	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
13240	209462	R. Bingham	England	Hamilton Academical FC	24	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
13241	219702	K. Dankowski	Poland	Slask Wroclaw	21	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
13242	225590	I. Colman	Argentina	Club Atlético Aldosivi	23	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
13243	233782	M. Feeney	England	Everton	19	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
13244	239158	R. Minor	Denmark	Hobro IK	30	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
13245	242998	Klauss	Brazil	HJK Helsinki	21	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

```
[172] 1 len(df.loc[df['HeadingAccuracy'].isnull()])
```

48

```
[173] 1 len(df.loc[df['HeadingAccuracy'] > 0 ]) + len(df.loc[df['HeadingAccuracy'].isnull()])
```

18207

Los valores nulos que hemos analizado son solamente 48, los cuales no alteran demasiado a nuestro dataset

Vemos el caso del jugador que es el único con valores nulos en las columnas que vamos a analizar

12619	229751	E. Ozuna	Argentina	Aldosivi	22	52.0	66.0	64.0	77.0	82.0	80.0	67.0	68.0	57.0	59.0
12792	233143	L. Sapetti	Argentina	Club Atlético Aldosivi	29	65.0	60.0	61.0	69.0	59.0	57.0	59.0	47.0	43.0	63.0
12829	215234	I. Quílez	Argentina	Club Atlético Aldosivi	28	49.0	63.0	58.0	71.0	81.0	70.0	38.0	49.0	59.0	62.0
13242	225590	I. Colman	Argentina	Club Atlético Aldosivi	23	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
13496	236667	J. Iritier	Argentina	Club Atlético Aldosivi	23	42.0	68.0	67.0	67.0	63.0	65.0	59.0	47.0	55.0	36.0
14859	232531	R. Garay	Argentina	Club Atlético Aldosivi	21	47.0	61.0	64.0	69.0	59.0	59.0	58.0	48.0	63.0	50.0
16005	243810	L. Ingolotti	Argentina	Club Atlético Aldosivi	18	12.0	10.0	12.0	33.0	39.0	28.0	37.0	17.0	45.0	20.0

Lo que se decidió fue colocarle valores de la media para no afectar el análisis.

```
[175] 1 df.fillna(df.mean(), inplace=True)
```

ahora todo el conjunto de valores NaN poseen los valores de la media de su respectiva columna

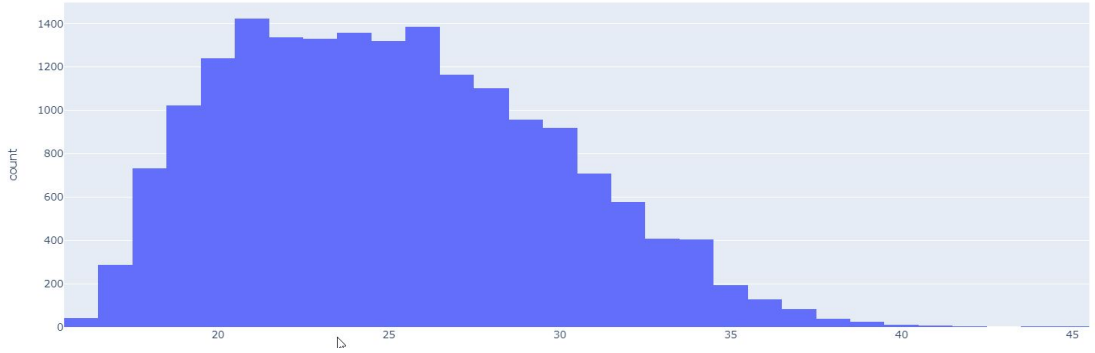
6. Graficos

6.1. Distribución de las columnas numéricas con relación a la media y la desviación estándar

Age:

```
La Media de la Edad:           25.122205745043114
Desviacion Standar de la Edad: 4.669814465849161
Mediana de  la Edad: 25.0
Moda de  la edad: 0    21
```

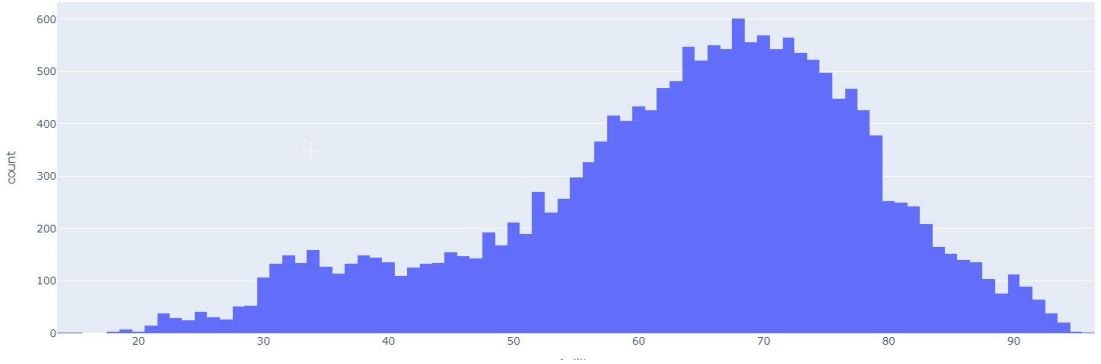
La Media de la Edad: 25.122205745043114
Desviacion Standar de la Edad: 4.669814465849161



Vemos que la edad media de los jugadores es de 25 con una distribución gausiana casi simétrica normal donde existen más datos entre los 20 y 30 años. Aqui podriamos aclarar que la edad del jugador es un tema muy importante ya que el tiempo de vida del jugador “profesional” llegaría terminar arriba de los 35 años

Agilidad

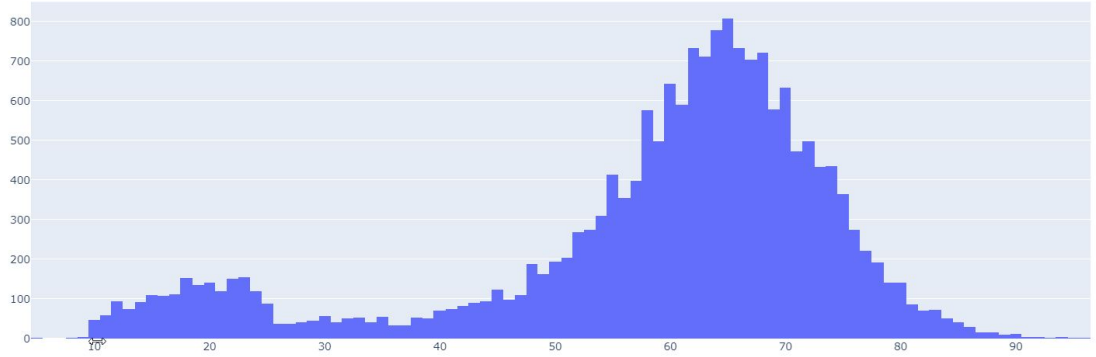
```
La Media de la Agilidad:           63.50360702681852
Desviacion Standar de la Agilidad: 14.746166139535365
Mediana de  la Agilidad: 66.0
Moda de  la Agilidad: 0    68.0
```



HAy una distribución gaussiana estándar porque se concentran los valores cerca de la media una distribución casi simétrica. Vemos outliers que están del valor de agilidad menores a 40. hay pocos jugadores con una agilidad fuera del promedio lo que la mayoria les podria llamar Jugadores de Elite. vemos que los existen varios datos dispersos debido a que la desviación estándar nos muestra un valor de ~15

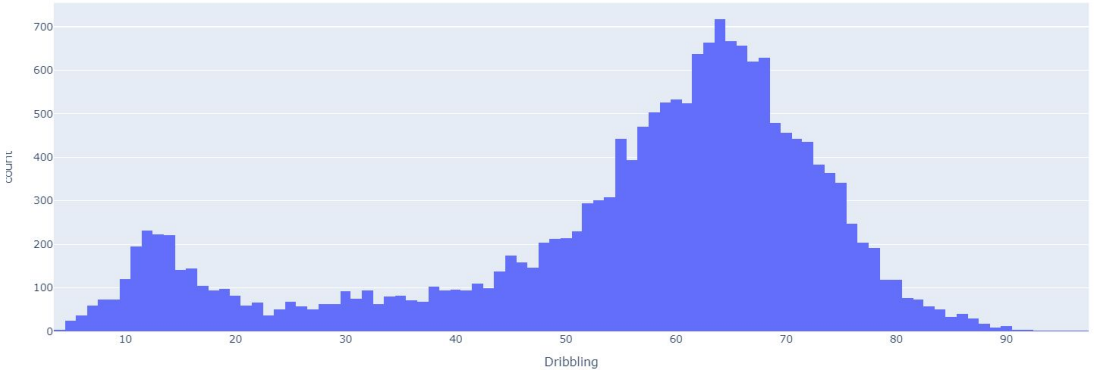
BallControl

```
La Media del Control del balon:          58.36945867063159
Desviacion Standar del Control del balon:16.66412618927394
Mediana del Control del Balon: 63.0
Moda del Control del Balon: 0    65.0
```



Vemos que esta habilidad tiene outliers para valores < a 40. y que la media está alejada de la campana gaussiana que determina el pico de los valores. Está distribución es asimetrica está alejada de la media ya que no forma la campana gaussiana por que el pico de valores está lejos de la media, además por los outliers que existen.

Dribbling

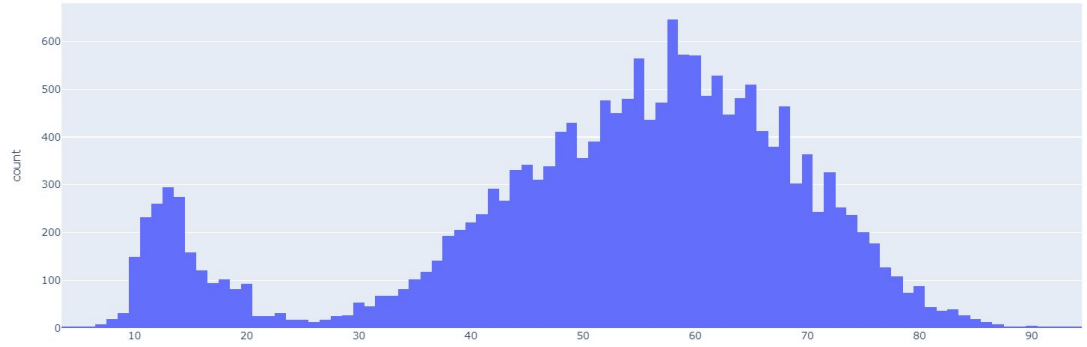



```
La Media de la Habilidad de movimientos y amagues: 55.37100060576021
Desviacion Standar de la Habilidad de movimientos y amagues: 18.884907473009555
Mediana de la Habilidad de movimientos y amagues: 61.0
Moda de la Habilidad de movimientos y amagues: 0    64.0
```

Vemos que esta habilidad tiene outliers para valores < a 45. y que la media está alejada de la campana gaussiana que determina el pico de los valores. Está distribución es asimetrica está alejada de la media ya que no forma la campana gaussiana por que el pico de valores está lejos de la media, además por los outliers que existen.

HeadingAccuracy

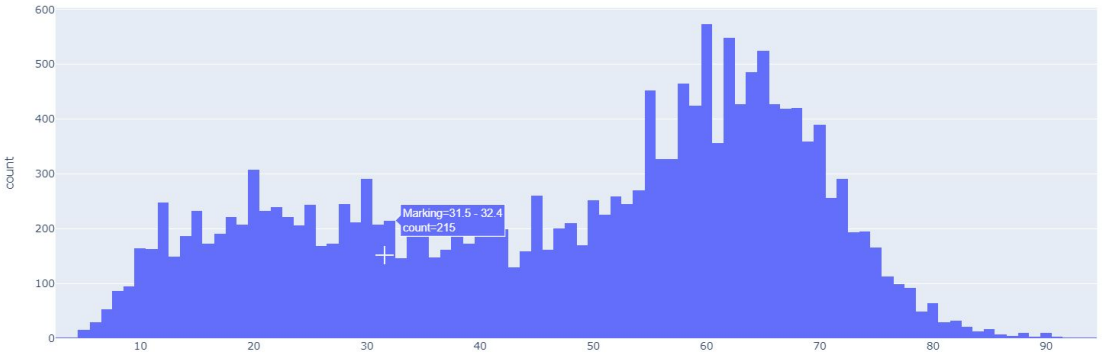
```
La Media de la Golpear con la Cabeza:           52.29814417093443
Desviacion Standar de  Golpear con la Cabeza: 17.356506062884993
Mediana de  Golpear con la Cabeza: 56.0
Moda de  Golpear con la Cabeza: 0    58.0
```



Vemos que esta habilidad es asimétrica, casi simétrica ya que los valores de la media, moda y mediana están levemente dispersos. existen outliers cuando la headingAccurracy tiene un valor menor a 30.
vemos que se forma la campana distribución gaussiana

Marking

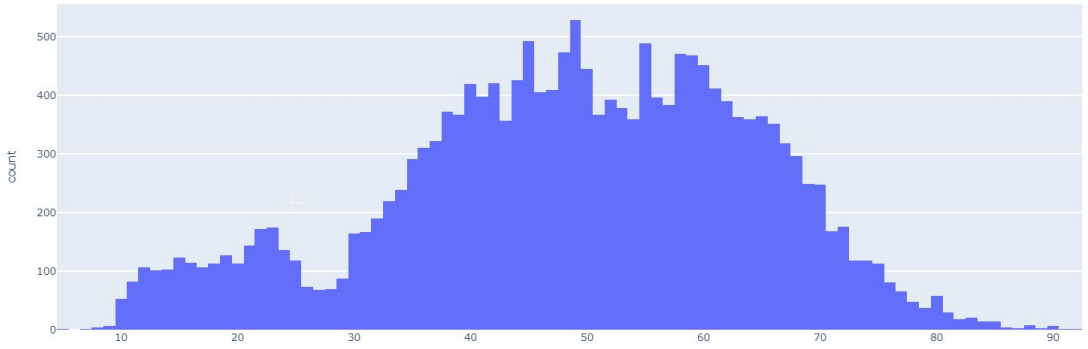
```
La Media de la Marcacion:           47.28162343741397
Desviacion Standar de la Marcacion: 19.87759513256158
Mediana de  Marcacion: 53.0
Moda de  Marcacion: 0    60.0
dtype: float64
```



Vemos que se forma una distribución segmentada. ya que existe un segmento entre valores de 10 y 40 y otro entre 50 y 80. vemos que la media está entre estos dos segmentos y la moda es 60 yéndose hacia el segundo segmento.
Esto podría darse ya que en el fútbol existen 2 especialidades bien marcadas que son los Defensores y los Delanteros. Los defensores tienen una alta habilidad para la marcación evitando que el oponente trate de convertir un gol. En cambio los Delanteros no tiene muy bien desarrollada esta habilidad. Un entrenador de fútbol no colocaría a un Delantero para defender pero si para anotar.

Penalties

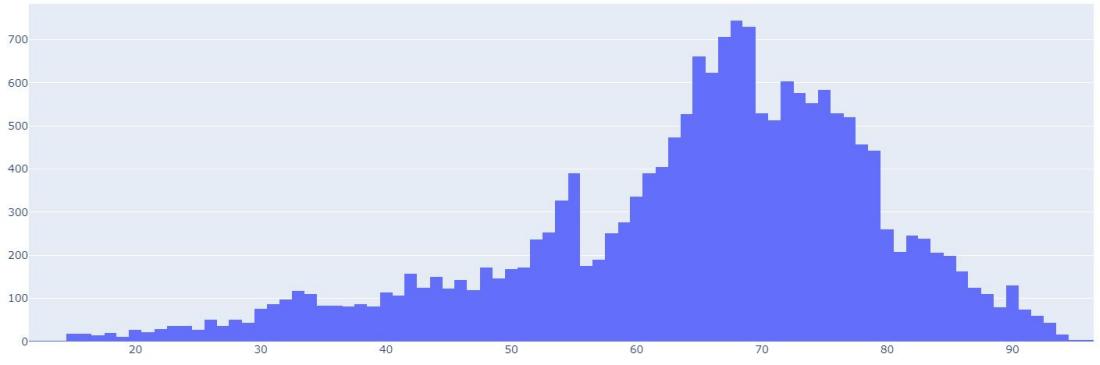
```
La Media de faltas:           48.548598491106375
Desviacion Standar de faltas: 15.682907309873777
Mediana de  faltas: 49.0
Moda de  faltas: 0    45.0
```



Vemos una distribución asimétrica con una alta tendencia a ser simétrica ya que la media moda y mediana están casi cerca. Existen outliers cuanto el valor de la habilidad es menor a 30.
Está columna indica las eficacia para convertir en gol una falta penal

SprintSpeed

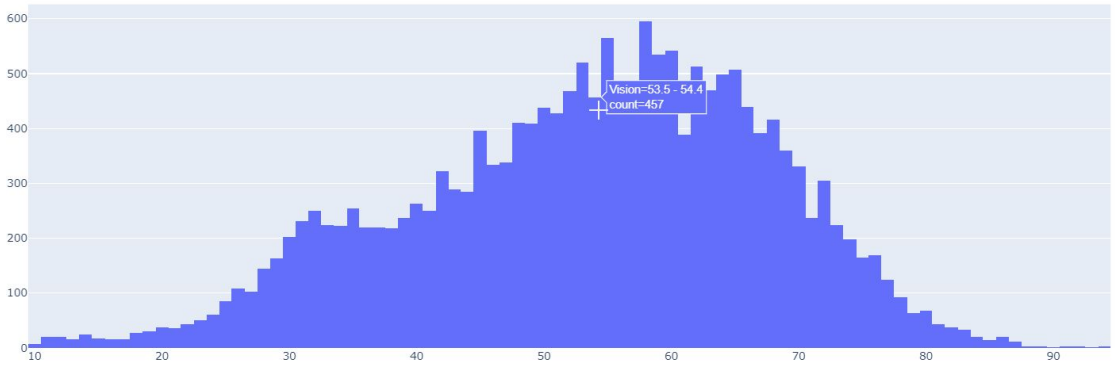
```
La Media de la SprintSpeed:         64.72696734401686
Desviacion Standar del SprintSpeed: 14.630226124525704
Mediana de  SprintSpeed: 67.0
Moda de  SprintSpeed: 0    68.0
```



Vemos una distribución asimétrica casi simétrica ya que la media, la moda y mediana no se acercan, vemos una diferencia pequeña. Hay una alta cantidad de jugadores que poseen la habilidad de Velocidad de Sprint, que en el fútbol es una habilidad de correr bastante rápido por un determinado tiempo con el balón

Vision

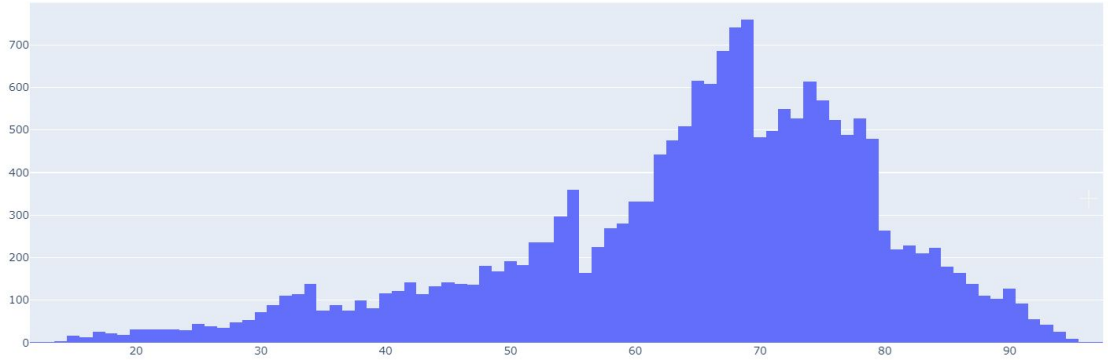
```
La Media de la Vision:           53.400903133432486
Desviacion Standar de la Vision: 14.12783134291396
Mediana de  Vision: 55.0
Moda de  Vision: 0    58.0
```



Está es la habilidad que tiene el jugador para ver el terreno de fútbol, ver a sus compañeros, al oponente. vemos que es una distribución asimétrica cercana a ser simétrica porq los valores de la media, moda y mediana están casi cerca. Se observan outliers de valores < a 25. y se observa un 2do pico entre 30 y 35

Acceleration

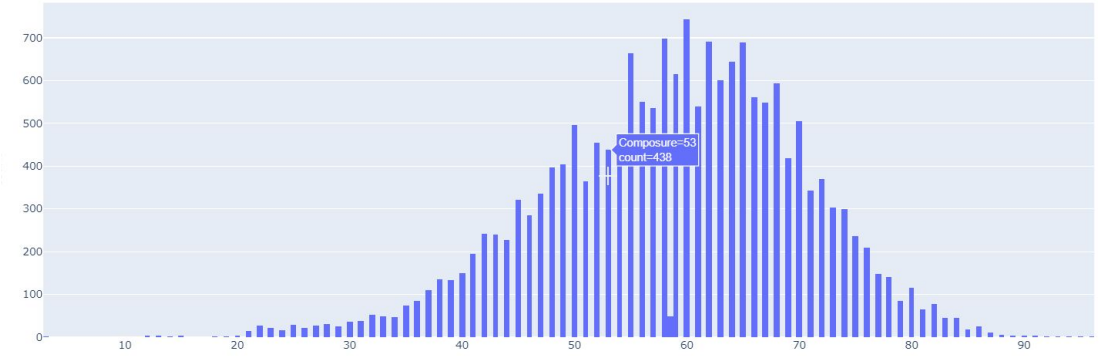
```
La Media de la Aceleracion:           64.61407566495967
Desviacion Standar de la Aceleracion: 14.907678842253612
Mediana de  Aceleracion: 67.0
Moda de  Aceleracion: 0    69.0
```



Se ve que existe una distribución casi simétrica ya que los valores de la media, moda y mediana no están muy alejados para formar una distribución simétrica. Vemos 4 picos entre 35, 55, 68 y 75 aproximadamente. Vemos que muchos jugadores tienen un valor de aceleración de 69 y está bien marcado los jugadores que tienen más de 80 de valor de aceleración

Composure

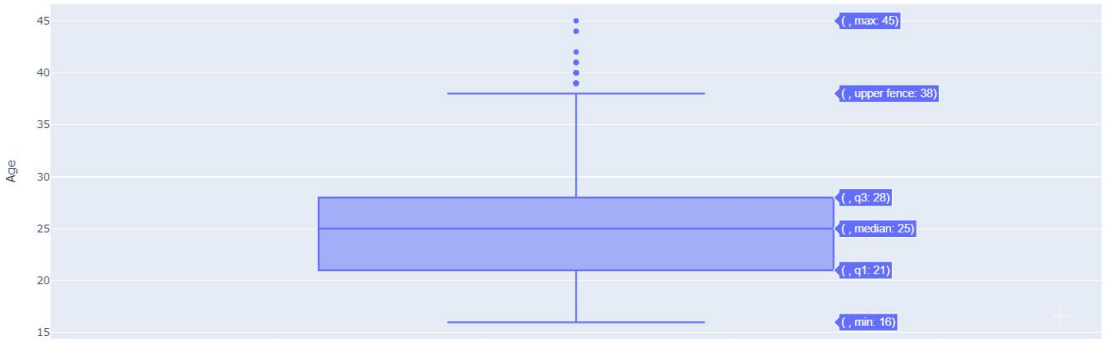
```
La Media de la Composure:            58.6482735833472
Desviacion Standar de la Composure: 11.420733726897799
Mediana de  Composure: 59.0
Moda de  Composure: 0    60.0
```



Está habilidad denota la eficacia del jugador dentro del campo de juego. Vemos que existe una distribución simétrica ya que la moda, media y mediana tiene valores casi exactos

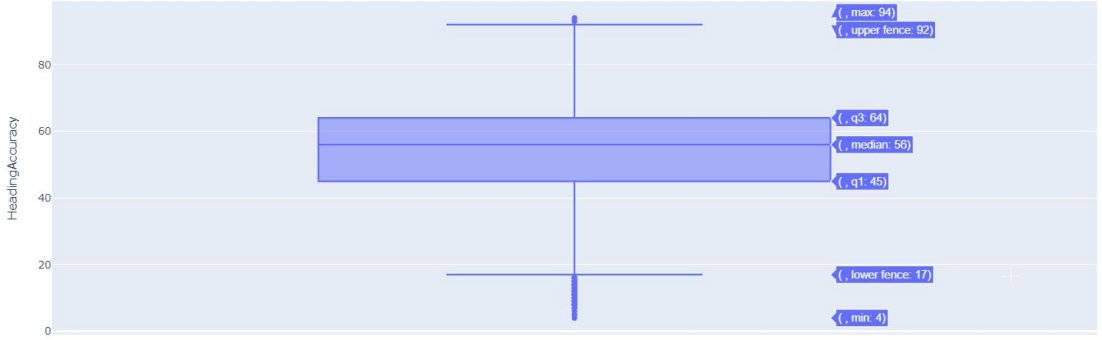
7. Analisis de Outliers

Age



Se observan outliers arriba de 38. Podemos indicar que hay jugadores profesionales contados por encima de esta edad hasta los 45 años. Vemos que existen jugadores ya profesionales que están jugando a partir de los 16 años y que el rango preferido es de 21 a 28.

HeadingAccurracy



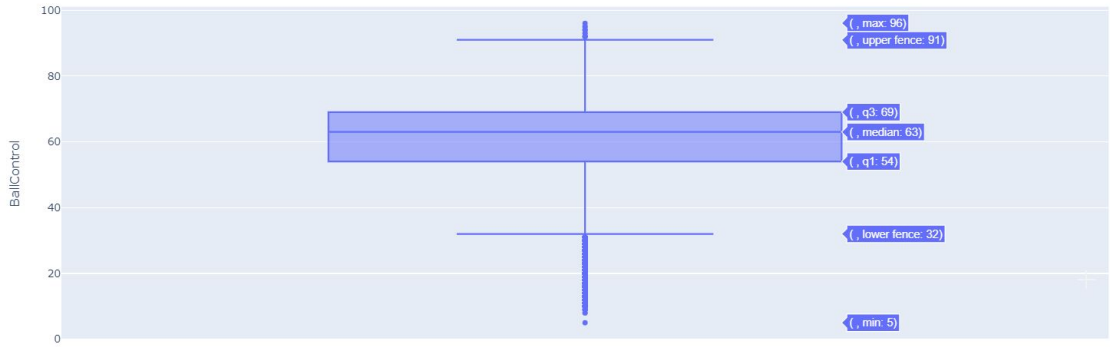
Se observan outliers en los extremos superior e inferior. Superior arriba de 92 e inferior debajo de 17.
Los outliers inferiores a 17 podríamos indicar que son los jugadores que se desempeñan como arqueros ya que ellos no trabajan esta habilidad

Dribbling

Se observan más outliers en la zona inferior por debajo de 21. en la zona superior por arriba de 96 se observan muy pocos

Los outliers inferiores a 21 podríamos indicar que son los jugadores que se desempeñan como arqueros ya que ellos no trabajan esta habilidad

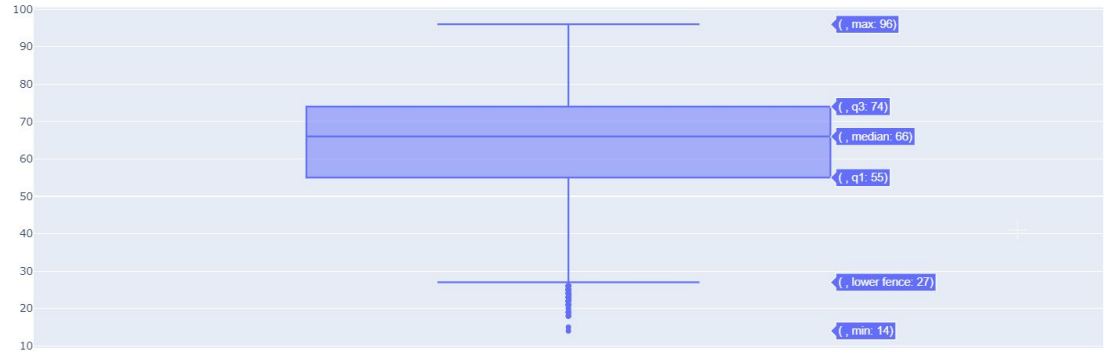
Ball Control



Se observa bastantes outliers en la zona inferior a 32 y Outliers en la zona superior arriba de 91

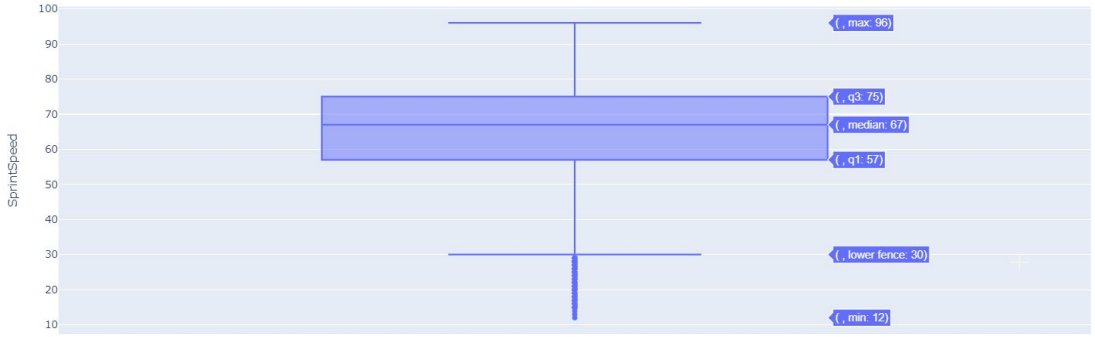
Los outliers inferiores a 32 podríamos indicar que son los jugadores que se desempeñan como arqueros y defensas ya que ellos no trabajan esta habilidad o prefieren trabajar en otra habilidad
Los outliers que están por encima de 91, vendrían a ser jugadores que son referentes del fútbol elite mundial como el caso de Lionel Messi, Cristiano Ronaldo, Neymar

Agility



Se observan outliers en la zona inferior a 27. No se observan outliers en la zona superior.
los Outliers de la zona inferior, podrían indicar a los jugadores con bajo desempeño en esta habilidad pero quizás tienen trabajada otra habilida.

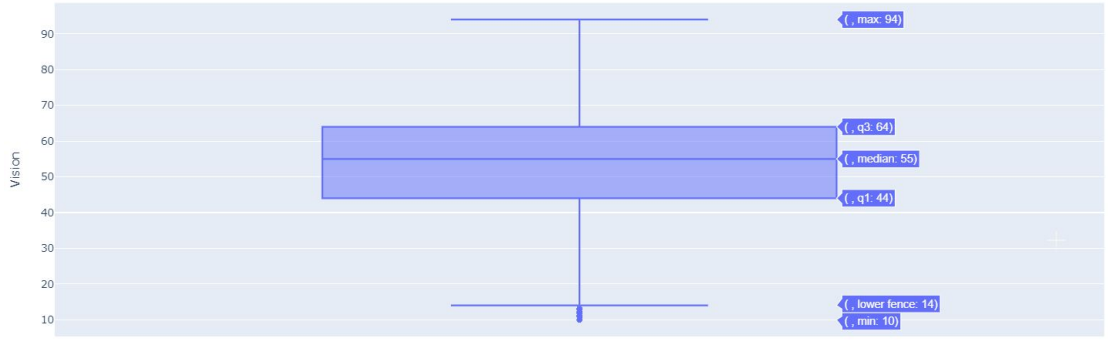
SprintSpeed



Se observan outliers en la zona inferior a 30 y no existen outliers en la zona superior.

Los outliers de la zona inferior, podrán establecer a jugadores que se desempeñan como arqueros ya que un arquero no trabaja ni desarrolla mucho esta habilidad.

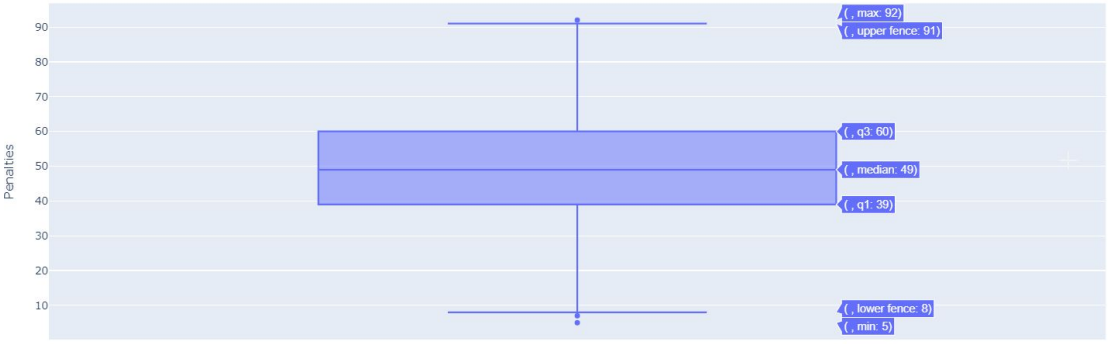
Vision



Se observan outliers en la zona inferior a 14.

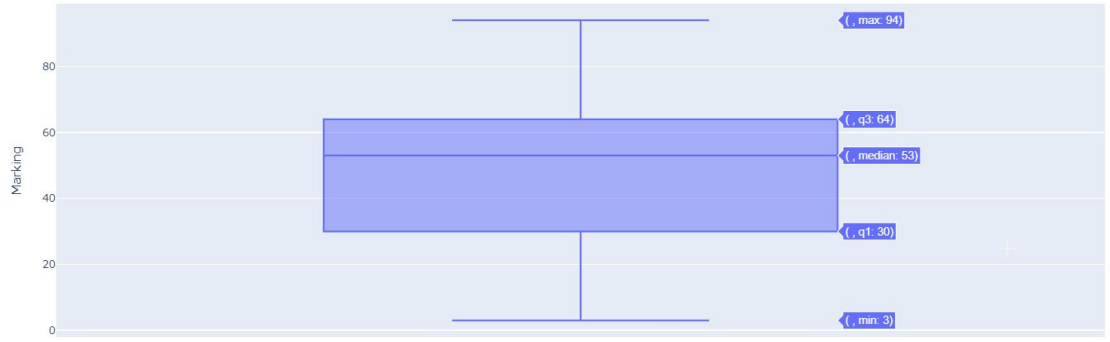
Se ven muy pocos ya que esta habilidad debe ser muy trabajada dentro de los jugadores en todas las posiciones

Penalties



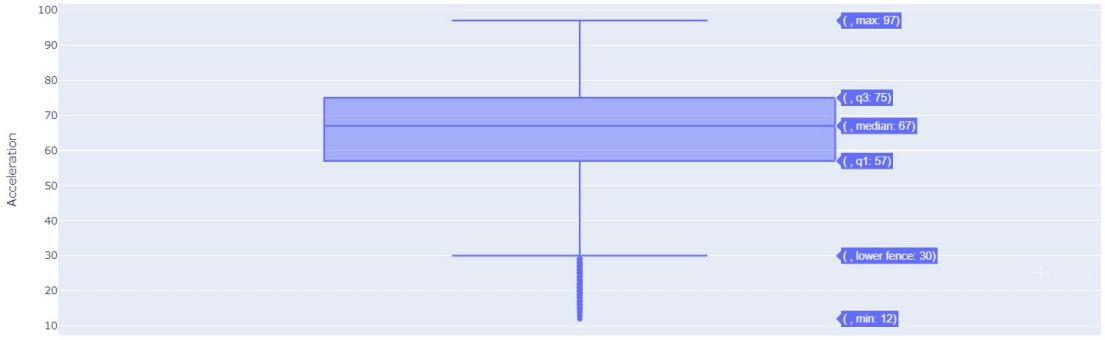
Se observan un outlier en la zona superior de 91 y en la zona inferior por debajo de 8
No influyen mucho estos valores ya que no afectaría en el cálculo deseado

Marking



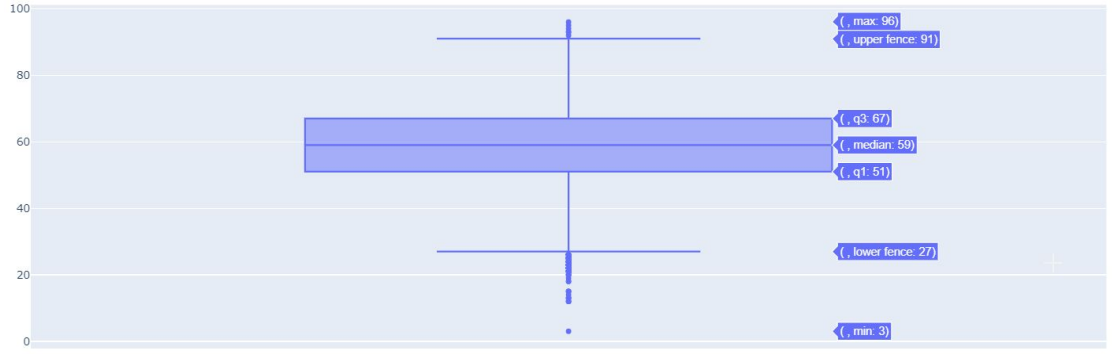
No se observan outliers, pero si se observa que la media está cerca del q3
Está es una habilidad bien trabajada. Existen jugadores que no trabajan esta habilidad pero no se observan como outliers, sino que están dentro de los parámetros establecidos

Acceleration



Se observan outliers en el limite inferior a 30.

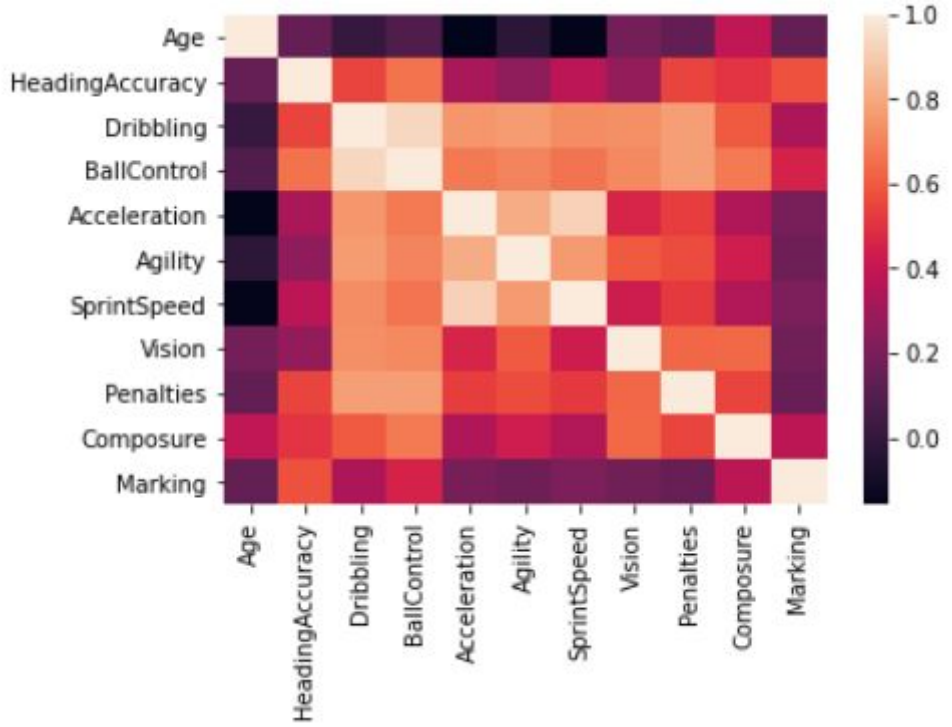
Composure



Se observa valores en el límite superior a 91 y en el límite inferior por debajo de 27.
los de arriba de 91 podrían ser los delanteros que tienen una eficacia al momento de convertir un gol
En el límite inferior se puede observar que estos outliers son jugadores que no son eficaces al momento de convertir o nunca han convertido una anotación o no tienen la posibilidad de convertir.

8. Calculo, análisis y gráfica de la correlación

	Age	HeadingAccuracy	Dribbling	BallControl	Acceleration	Agility	SprintSpeed	Vision	Penalties	Composure	Marking
Age	1.000000	0.147009	0.010154	0.084868	-0.158479	-0.019372	-0.151502	0.187200	0.139370	0.390560	0.142647
HeadingAccuracy	0.147009	1.000000	0.550750	0.658175	0.329647	0.260514	0.379453	0.275673	0.551978	0.507208	0.583123
Dribbling	0.010154	0.550750	1.000000	0.938942	0.748292	0.765153	0.726835	0.730150	0.769594	0.597498	0.336072
BallControl	0.084868	0.658175	0.938942	1.000000	0.675737	0.704604	0.663990	0.718411	0.769791	0.674881	0.452705
Acceleration	-0.158479	0.329647	0.748292	0.675737	1.000000	0.810832	0.921928	0.461552	0.532908	0.347427	0.195369
Agility	-0.019372	0.260514	0.765153	0.704604	0.810832	1.000000	0.763623	0.597327	0.566175	0.432511	0.167122
SprintSpeed	-0.151502	0.379453	0.726835	0.663990	0.921928	0.763623	1.000000	0.429554	0.521071	0.351607	0.212575
Vision	0.187200	0.275673	0.730150	0.718411	0.461552	0.597327	0.429554	1.000000	0.632927	0.636280	0.176760
Penalties	0.139370	0.551978	0.769594	0.769791	0.532908	0.566175	0.521071	0.632927	1.000000	0.551801	0.152296
Composure	0.390560	0.507208	0.597498	0.674881	0.347427	0.432511	0.351607	0.636280	0.551801	1.000000	0.384081
Marking	0.142647	0.583123	0.336072	0.452705	0.195369	0.167122	0.212575	0.176760	0.152296	0.384081	1.000000



Podemos observar que los siguientes valores están altamente correlacionados. Hemo tomado el valor mínimo para la correlación de 0.75

```
Dribbling        BallControl        0.938942
Acceleration     SprintSpeed        0.921928
                 Agility            0.810832
Penalties        BallControl        0.769791
Dribbling        Penalties          0.769594
                 Agility            0.765153
SprintSpeed      Agility            0.763623
```

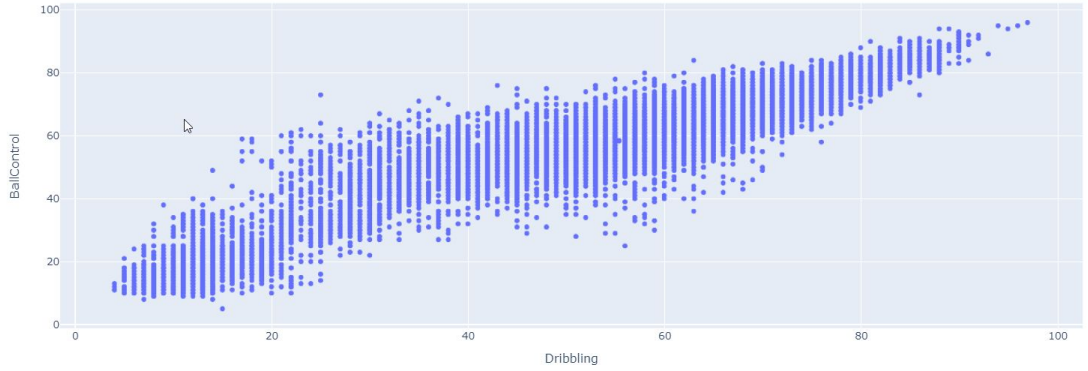
Vemos el Dribbling y el control del balón tienen correlación ya que si un jugador posee la habilidad del dribbling que es la de evadir al oponente y como el juego se desarrolla con el un balón, obviamente al tener un buen control del balón existe una alta probabilidad de eludir al oponente.

Vemos también que la Aceleración con la Agilidad, Velocidad de Sprint están relacionados si es Ágil y veloz en tramos cortos, el jugador tendrá fuerza interna para lograr una alta aceleración.

Vemos también que el Dribbling con la Agilidad están relacionados. ya que un jugador para eludir al oponente debe tener mucha agilidad.

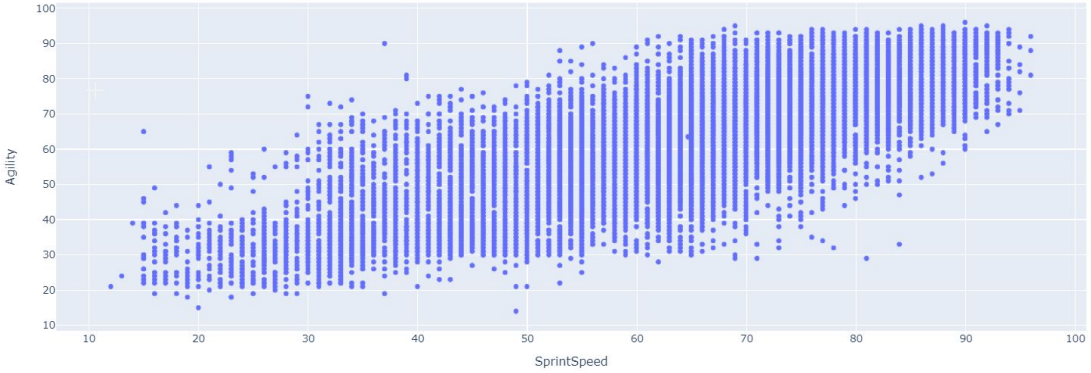
9. Comparación de variables del dataset *y* Scatterplots

Comparacion entre BallControl y Dribbling



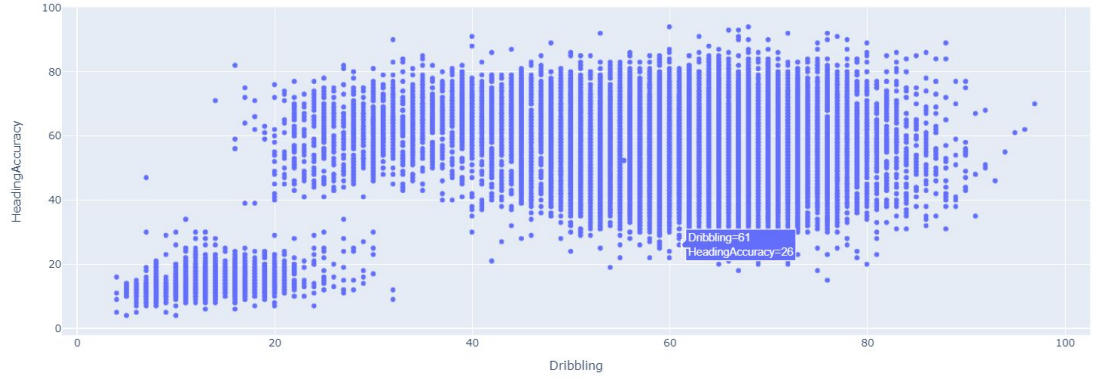
Aquí podemos ver la correlación que existe entre BallControl y Dribbling. Vemos que existe una alta correlación positiva lineal describiendo un patrón ascendente. Ya que a medida que incrementa X, incrementa Y

Agility y SprintSpeed



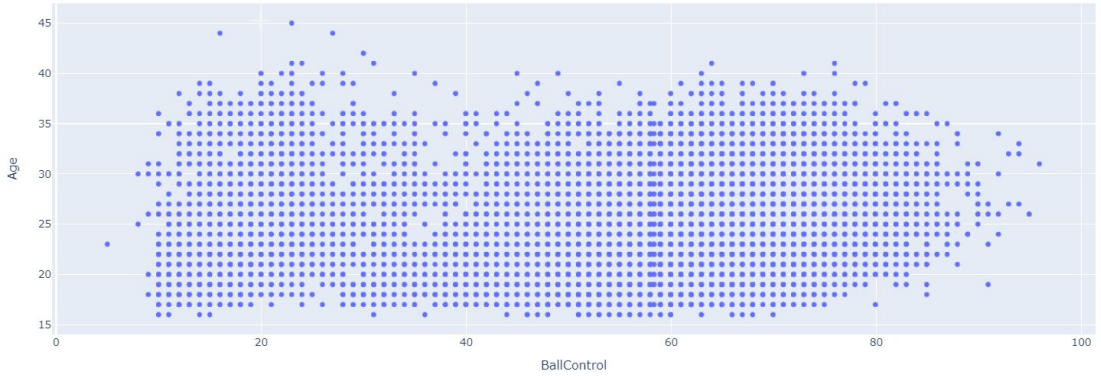
Aquí podemos ver la correlación que existe entre BallControl y Dribbling. Vemos que existe una baja correlación positiva lineal describiendo un patrón ascendente aunque están un poco más dispersos que el anterior análisis. Ya que a medida que incrementa X, incrementa Y

Dribbling y Heading Accuracy



Aquí podemos ver que no existe una relación ya que se forman 2 segmentos que no siguen un patrón determinado.

Ballcontrol y Age



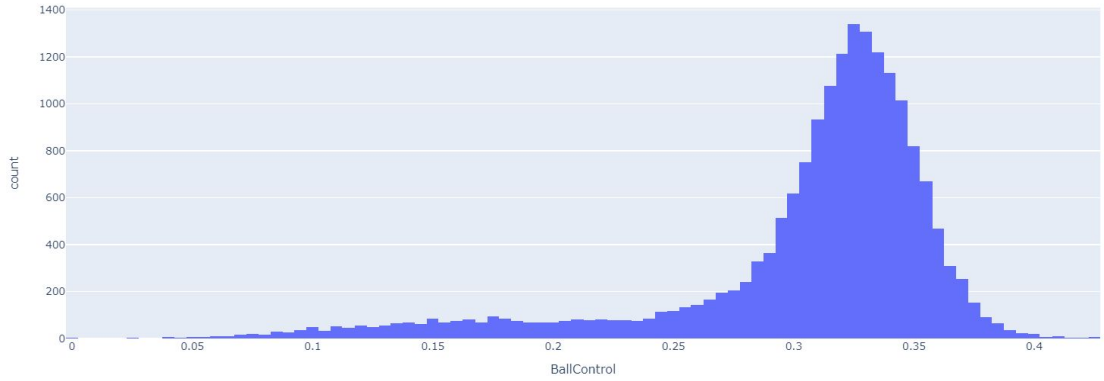
Aquí podemos ver que no existe una relación. No se apreció ninguna correlación entre las dos variables.

10. Normalización del dataset - Graficación de la distribución

Los valores que se tomaron para la normalización son de una escala de valor de 0 a 5

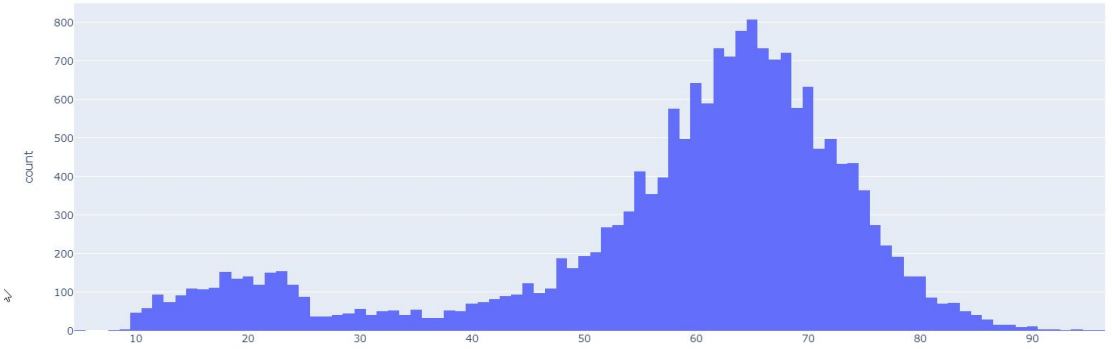
Normalizado - BallControl

```
1 px.histogram(df_normalized, 'BallControl')
```



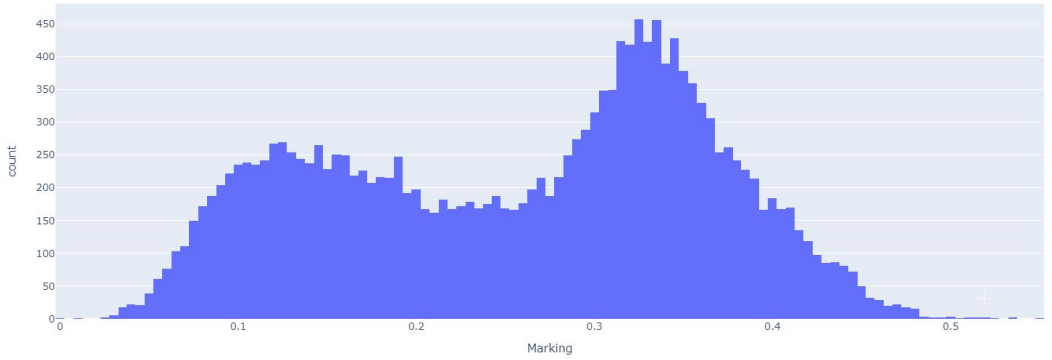
Original - BallControl

```
1 px.histogram(df, 'BallControl')
```



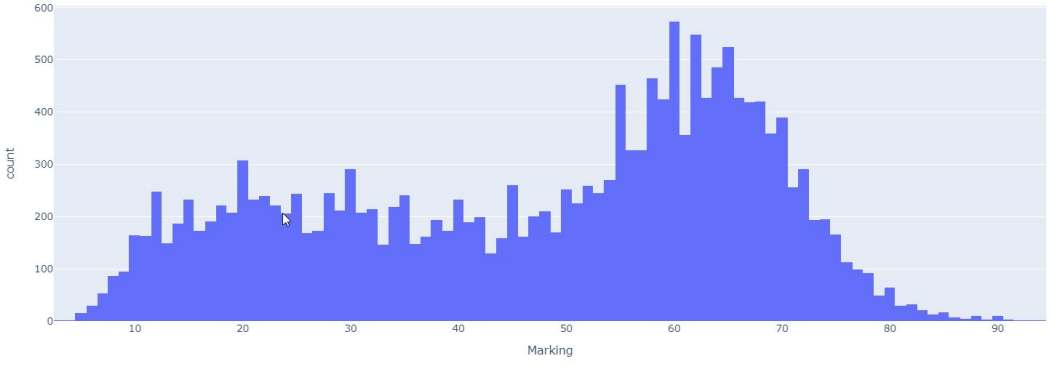
Normalizado - Marking

```
[252] 1 px.histogram(df_normalized, 'Marking')
```



Original- Marking

```
[253] 1 px.histogram(df, 'Marking')
```



11. Conclusiones

Hemos visto que mediante la correlación podríamos llegar encontrar las habilidades en las cuales podremos trabajar más en algunos jugadores tomando en cuenta la alta correlatividad que tienen las variables por ejemplo el Dribbling y el Ball Control, sabemos que estas están altamente correlacionadas y podríamos utilizarlas para tomar en cuenta en la contratación de un nuevo jugador. Esta habilidad podría determinar la posición en la cual un posible jugador o un jugador actual se desempeñaría mejor.

Hemos también visto que la Edad no está relacionada con la Aceleración, Agilidad y Velocidad de Sprint, esto significa que el jugador no depende mucho de la edad para ser ágil y veloz, y eso se puede ver en bastante casos. Obviamente el rendimiento del jugador no es el mismo, pero esta variable no ha sido tomado en cuenta.

Hemos visto que la Aceleración está altamente relacionada con la velocidad de Sprint y el Sprint con el Control de Balón y el Dribbling. Estas variables hacen tomar en cuenta para colocar a jugador en un juego que se necesitan un rápido desempeño y generación de juego