|  |  |  |
| --- | --- | --- |
| **Nombre**: Diego Ronaldo Sanchez Pichardo | | **Matrícula**: 05077303 |
| **Nombre del curso:** Ciencia de datos | **Nombre del profesor**: Kristel Cano Galán | |
| **Módulo**: 2do | **Actividad**: Actividad 2 | |
| **Fecha**: 31 de Marzo del 2025 | | |
| **Bibliografía**:  Pandas Development Team. (2024, September 20). pandas 2.2.3 [Software]. PyPI. <https://pypi.org/project/pandas/>  MongoDB, Inc. (n.d.). PyMongo: Python driver for MongoDB (Version [versión actual]). PyPI. <https://pypi.org/project/pymongo/>  MongoDB, Inc. (n.d.). Cree una base de datos Python con MongoDB. MongoDB. <https://www.mongodb.com/es/resources/languages/python> | | |

**Objetivo**

En esta actividad, manejaremos datos generados de manera aleatoria, sobre jugadores de basquetball. Nuestro objetivo sera generar interpretaciones sobre nuestros datos. Haremos un analisis previo al que nos servira para generar la interpretacion, donde conoceremos si hay valores atipicos, nulos o faltantes. Tambien, manejaremos interpretaciones textuales y en graficas, para tener un mejor entedimiento de un usuario no especializado en la ciencia de datos.

Para esta actividad usaremos Python junto a las siguientes librerias:

* Pandas: Manejo de datos y generacion de frames.
* PyMongo: Librería de MongoDB para manejar datos de una colleccion, gracias a la conexión con localhost.
* Faker: Generacion de datos aleatorios.
* Matplotlib: Generacion de graficas.
* Seaborn: Personalizacion de graficas.

**Preguntas clave**

Preparacion de los datos:

* ¿Hay datos nulos o faltantes en nuestra base de datos?
* ¿Cuántos jugadores juegan en cada posicion?
* ¿Qué equipos tienen mas jugadores?, ¿Cuántos jugadores juegan por equipo?

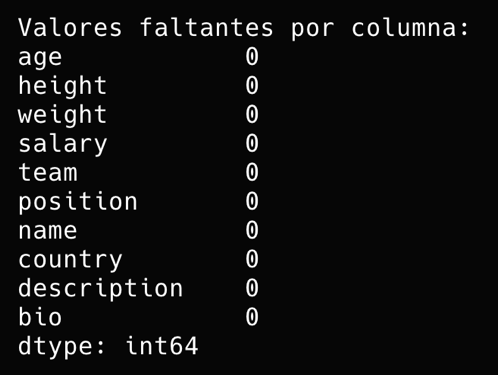
Graficas:

* ¿Cuántos jugadores hay en cada rangos de edad?
* ¿Cómo se distribuye la altura de los jugadores?
* ¿Cómo se distribuye la altura, según el peso de cada jugador?

**Preparacion de los datos**

Se realiza una analisis rapido en los datos, para poder conocer si faltan datos, hay datos nulos o atipicos en nuestra colleccion de datos. Esto se hace apartir del Data Frame generado con la funcion “DataFrame” la librería “pandas”, y usando la funcion de la clase, .isnull() para iterar cada dato y obtener los valores nulos y .sum() para sumar el conteo de datos nulos.

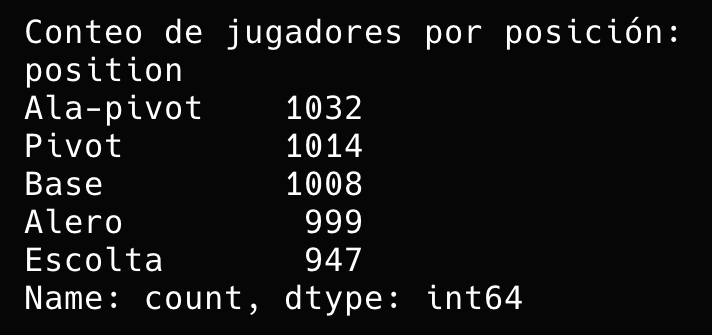
En nuestra colleccion no se obtienen datos nulos ni atipicos, y se genera la interpretacion por columna.



**Analisis exploratorio de datos**

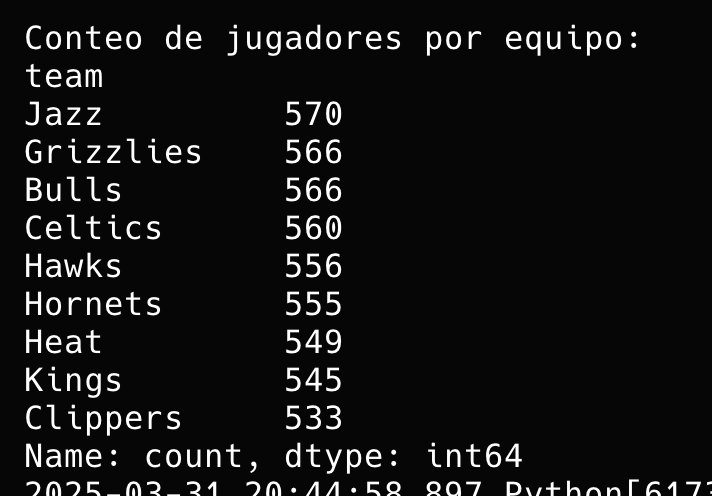
1. Conteo de jugadores por posicion

De manera rapida, podemos observar que la distribucion entre posiciones es muy cercana a su mediana. Podemos concluir que la mayoria de los jugadores juegan en la posicion del “Ala-pivot”, esto se puede deber a varios motivos como su altura, su peso, su estilo de juego, etc.



1. Conteo de jugadores por equipo

De misma manera que el anterior, podemos observar de manera rapida, que la distribucion de jugadores por equipo, es muy cercana a su mediana. Debemos de tener en cuenta que esto es una representacion, en la vida real, los equipos no poseen tantos jugadores. Los Jazz lideran con 570 jugadores, y se encuentran muy cerca de los Clippers con 533 jugadores.



**Graficas para visualizacion de datos**

1. Distribucion de Edad: Histograma

Podemos observar que en la mayoria de los rangos de edad de entre 0 a 250 se mantienen en la media de la grafica. Encontramos que hay mas de 400 jugadores en la minima, la maxima, y la mediana de edad en nuestra grafica. Hay más jugadores en los extremos (18-20 años y 39-40 años), lo que sugiere que hay una combinación de jóvenes talentos y jugadores veteranos. La línea KDE muestra fluctuaciones, lo que indica que las edades no siguen una distribución normal.

1. Diagrama de Cajas: Diagrama de cajas

La mediana (línea dentro de la caja) está cerca de los 200 cm, lo que indica que la mayoría de los jugadores miden alrededor de esa altura. La caja representa el rango intercuartil, con la mayoría de los jugadores entre 190 cm y 210 cm. No hay valores atípicos visibles, lo que indica que todas las alturas están dentro de un rango esperado.

1. Altura vs Peso: Grafica de dispersion

Se observa un patrón horizontal homogéneo, lo que indica que no hay una correlación clara entre altura y peso. Hay jugadores con una amplia variabilidad de peso incluso para la misma altura. Los valores parecen bien distribuidos sin anomalías significativas.

