**Predicción del Torneo de Baloncesto NCAA 2017**

**Integrantes del grupo:**

Andrés Darío Peña

Andrés Ramirez Aristizabal

Leider Caicedo Palacios

**Materia:**

Modelos y Simulación 1

**Docente:**

Raúl Ramos Pollán

Universidad de Antioquia

Facultad de ingeniería

Octubre 2023

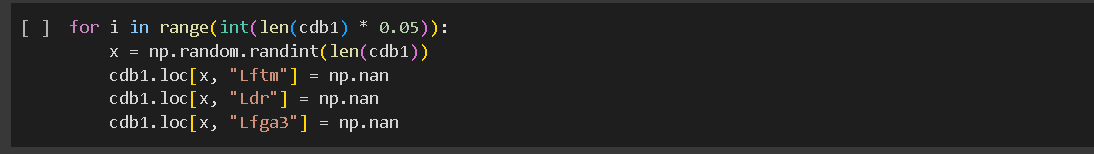
**Planteamiento del problema.**

Predecir los resultados de un torneo de baloncesto universitario masculino en Estados Unidos en el año 2017. Utilizaremos los resultados recopilados de torneos pasados para poder construir y probar diferentes modelos de predicción. Después pronosticamos los resultados de todos los enfrentamientos posibles en el torneo de 2017.

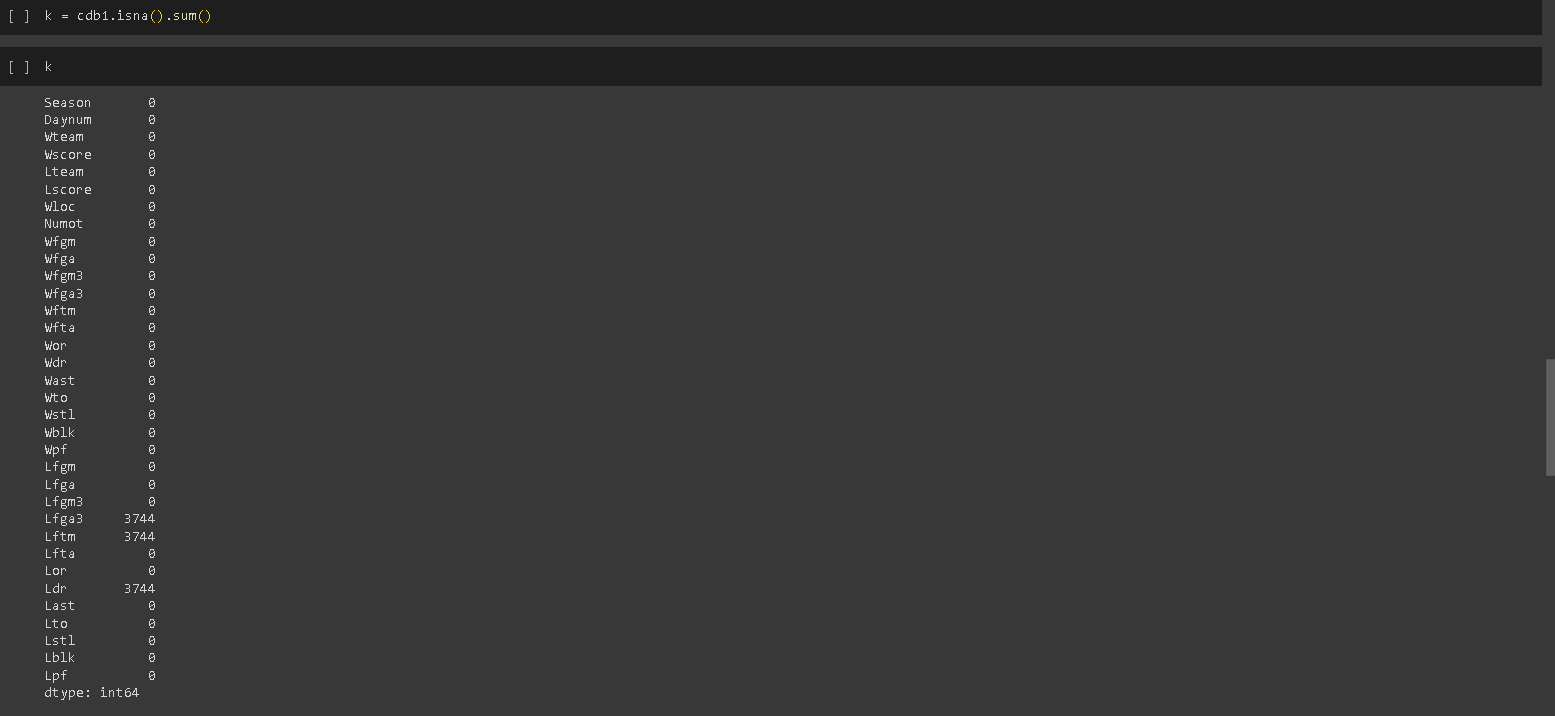
**Simulación del dataset para requerimientos:**

Se realizaron las simulaciones de los requerimientos pedidos para la base de datos seleccionada, los cuales fueron:

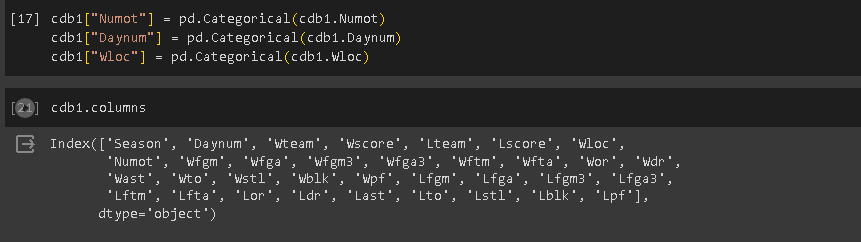
**El 5% de los datos de al menos 3 columnas deben ser faltantes:** Con este ciclo se ejecutó dicha simulación de datos faltantes en las columnas Lftm, Ldr y Lfga3

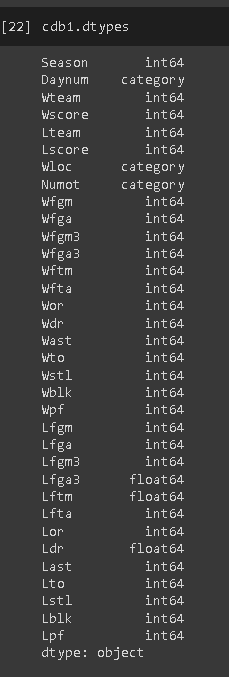


Este es el resultado presenciado donde el 3% de los datos de las columnas mencionadas (3744) se presentan como nulos gracias a la información proporcionada por la función isna de pandas dataframe.



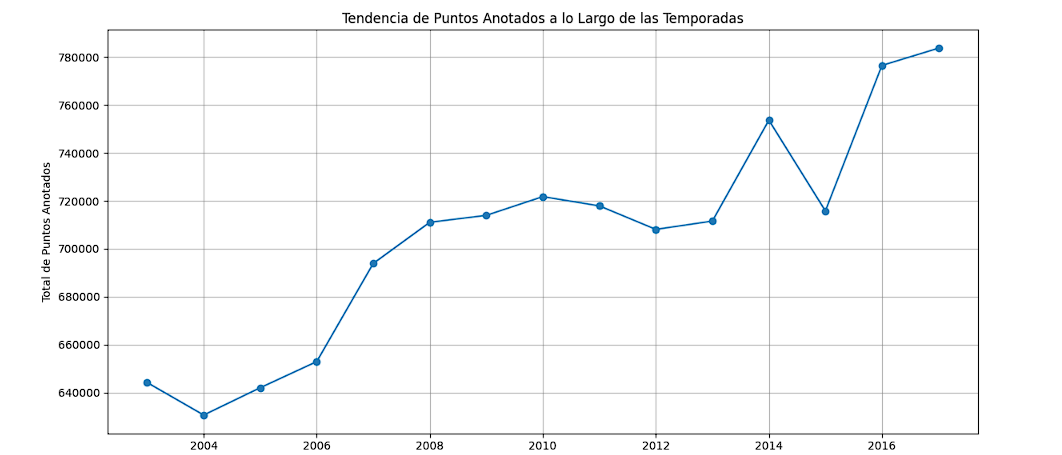
**Al menos un 10% de las columnas han de ser categóricas:** El 10% de las 34 columnas pertenecientes al data frame equivalen a 3 columnas redondeadas a entero (3,4 originalmente). Las columnas simuladas a ser establecidas como categóricas han sido: “Numot”, “Daynum” y “Wloc”.



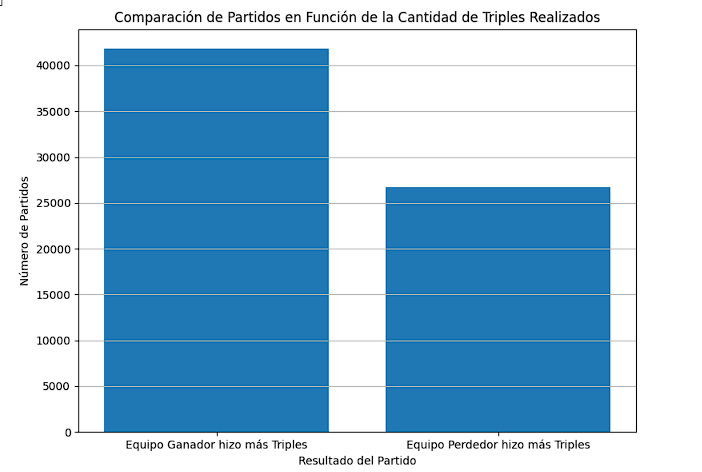


**Exploración de datos.** Inicialmente tomamos el conjunto de datos que seleccionamos para analizarlo, revisamos las distintas variables que lo conforman, los datos asociados a estas variables y la relación entre estos datos, todo esto para definir nuestro modelo de forma eficiente y acertada.

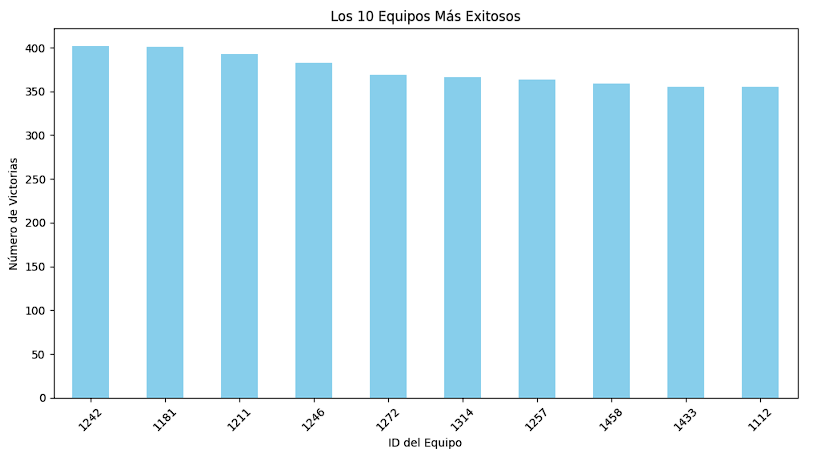
**Gráficos obtenidos.** Con el análisis y el proceso anterior realizado en los datos hemos obtenido algunos gráficos en los que se reflejan estadísticas comparativas entre partidos como la cantidad de puntos anotados, los triples anotados, faltas, robos, etc. También se observan otras estadísticas relacionadas a los resultados de los encuentros.

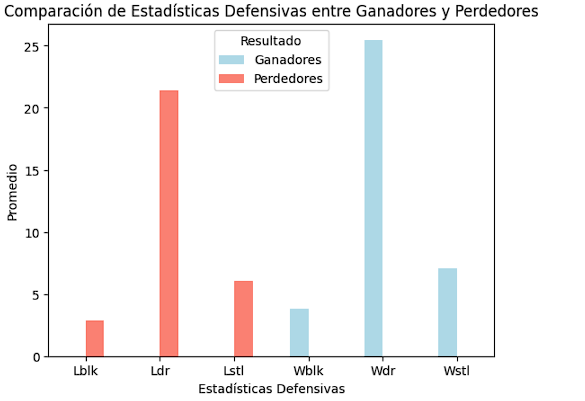


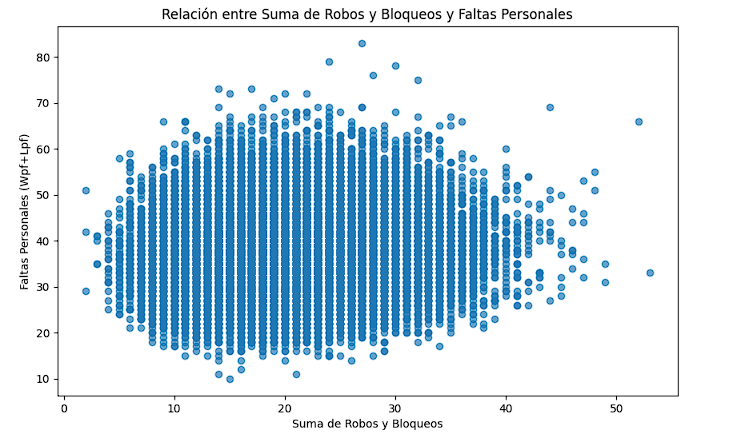
Observaciones: A medida que transcurren los años el baloncesto se vuelve un deporte más ofensivo, un deporte en el que se convierten cada vez más puntos , aunque existe una caída en los puntos en el año 2015 en comparación con el año anterior(2014).



Este gráfico muestra que los equipos ganadores en general convierten más triples que los equipos perdedores.

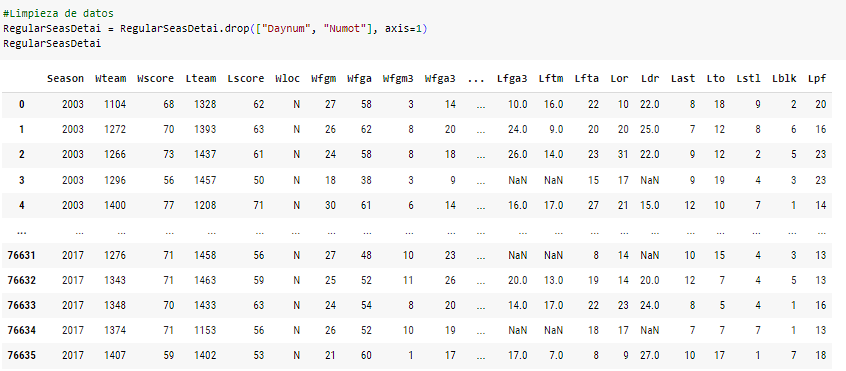






En este gráfico de dispersión, los bloqueos (wblk+lblq) más los robos (wstl+lstl) se representan en el eje x, y las faltas personales (wpf+lpf) en el eje y. Cada punto en el gráfico representa un partido ganado por un equipo. Si se observa una tendencia en la que equipos con más faltas personales también tienen más robos y bloqueos, podría sugerir que están dispuestos a asumir más riesgos en defensa para intentar obtener ventajas en el juego.

Se realiza la **limpieza de datos** al eliminar las columnas innecesarias “Daynum” y “Numot” las cuales contienen el día en que se jugó el partido y Numot el cual indica el número de períodos de tiempo extra en el juego.



Este gráfico te ayuda a visualizar la relación entre estas variables y evaluar si existe una correlación entre las faltas personales, los robos y los bloqueos en los equipos ganadores.

**Preprocesado de datos.** Después realizaremos un preprocesado de datos, limpiando y puliendo un poco el dataset, estableciendo un conjunto de datos sólido y completo, teniendo en cuenta la usabilidad de variables categóricas, la eliminación de columnas innecesarias y el filtrado de datos valiosos para las predicciones; entre otras cosas a tomar en cuenta.