ENTREGA 2

Proyecto NCCAA's clairvoyance.

Intengrantes

Oriana Mejía Cardona – CC. 1035879334 – Ing. Mecánica

PROGRESO ALCANZADO

Se llevaron a cabo 2 procesos que están en diferentes notebooks, ya que se iban creando archivos nuevos que se utilizarán en los siguientes notebooks; los datasets empleados están guardados en github y de allí sin cargados al colab, sin necesidad de descargarlos en la maquina local, de esta manera los notebooks se pueden reproducir sin problema. A continuación, se describen los 2 procesos realizados:

Simulación de datos

En el archivo 01 – Simulación de datos.ipynb se muestra el dataset original (76636 x 34), este contaba solo con una variable categórica por lo que fue necesario hacer un ciclo para convertir variables numéricas a categóricas.

```
def column_to_discrete(column_df):
    for i in range(len(column_df)):
        if column_df[i]<=33:
            column_df.loc[i] = "L"
        elif column_df[i] >=66 :
            column_df.loc[i] = "H"
        else:
            column_df.loc[i] = "M"
```

Esto se llevo a cabo con 4 columnas con valores entre 0 y 100. De esta forma se cumple con el requisito de que al menos 10% de las columnas sean categóricas.

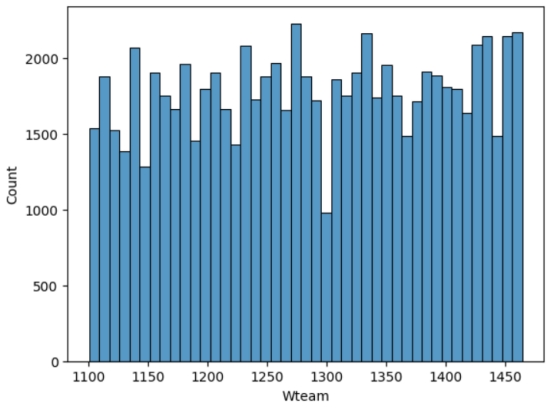
Este data set tampoco tenía valores nulos por lo que se escogieron 3 columnas y de manera aleatoria se eliminaron 5% se sus datos.

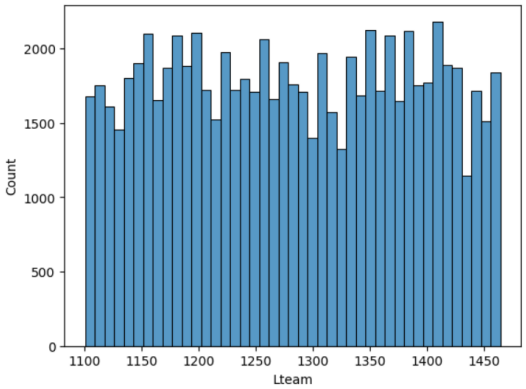
```
for i in range(int(len(df) * 0.05)):
    x = np.random.randint(len(df))
    df.loc[x, "Lftm"] = np.nan
    df.loc[x, "Ldr"] = np.nan
    df.loc[x, "Lfga3"] = np.nan

df
```

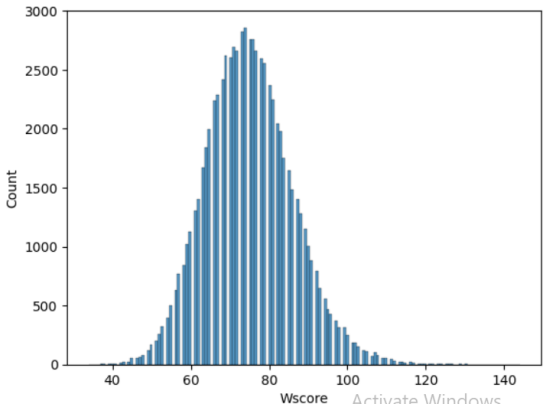
Exploración de datos

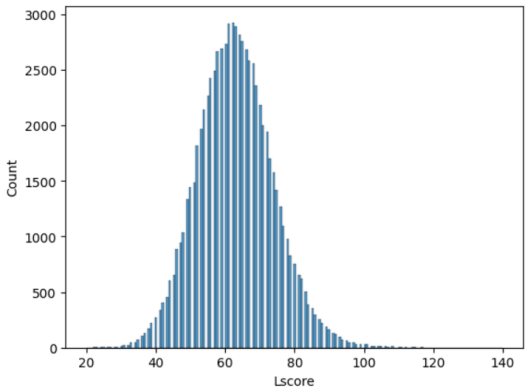
Después de realizar la simulación de los datos se obtiene un archivo llamado *RegularSeasonDetailedResults_processed.csv,* con el que realizaremos la exploración de datos para identificar patrones, graficarlos y conocer los datos.

- **Tamaño del dataset:** El tamaño del dataset es de (76636 , 34)
- **Valores nulos:** La columnas Lftm, Ldr y Lfga3 tiene 76636 valores nulos (aquí podemos encontrar un error que no pude corregir y que fue un problema durante el desarrollo del notebook)
- **Variables objetivo:** Como no tengo mucha claridad con el tema, no he podido escoger exactamente con cuál variable trabajaré como objetivo, las seleccionadas son: Wteam, Lteam, Wscore, Lscore.









- **Tipos de datos:** Los tipos de datos son int64 y float64 en gran mayoría y otros object en las variables que se convirtieron en categóricas-
- **Inspección de datos numéricos:** En esta inspección se calcula la desviación estándar, valor mínimo, máximo, percentiles de las columnas numéricas y promedio.

- **Inspección de variables categóricas:** Quería hacer una inspección detallada de las variables que se convirtieron a categóricas pero no tuve las completas claridades para ejecutar los histogramas para encontrar una relación entre las variables categóricas y las variables de interés escogidas.

Conclusiones

Por ahora todavía se necesita hacer más exploración de datos y definir claramente cual será la variable objetivo, además creo que será necesario combinar datos de otros datasets proporcionados por la competencia de Kaggle para tener una predicción más acertada de quien será el posible campeón de la siguiente temporada. Por otro lado me cuesta un poco la manipulación de los datos por lo que me ha sido más difícil poner practica lo aprendido en el curso.