Memoria sobre EDA: Estudio comparativo para analizar si Omar Vizquel tiene sustento estadístico para opositar al Salón de la Fama

Antecedentes

El beisbol es uno de los deportes que más me gustan, en especial por la importancia que cobran las estadísticas en el juego, de allí que me haya decantado por esta temática para el análisis exploratorio de datos.

De las muchas posibles líneas para analizar, me decanté por revisar los números de Omar Vizquel, un jugador que no goza de gran popularidad, pero que al chequear sus números, indudablemente se puede debatir su pertenencia al Salón de la Fama del béisbol norteamericano, sobre todo, si se comparan con los de Ozzie Smith, un jugador de la misma posición, y habilidades similares que entró al Salón de la Fama en su primera boleta, casi de forma unánime, y que si gozaba de una gran popularidad en su momento, al que además, comúnmente se le reconoce como el mejor campocorto de la historia.

Haciendo el EDA

Como fuente de datos, utilicé la web https://www.baseball-reference.com/ que cuenta con la información bastante estandarizada, y para la extracción, lo hice de 2 formas:

- Usando web scrapping para las estadísticas ofensivas. Podía haberlas descargado en Excel simplemente, pero quise dotar al trabajo de un toque de calidad.
- Para las estadísticas defensivas, no me fue posible implementar web scrapping así que las descargué en Excel y las trabajé con read_excel en el Notebook.

Al tratarse de 2 jugadores defensivos, las estadísticas de fildeo eran necesarias, y, siendo que se trata de un análisis para el ingreso al Salón de la Fama, existe un sesgo importante por el apartado ofensivo.

Para la limpieza de los datos, al conocer del tema, sabía que métricas utilizaría. Usé .iloc para limpiar las filas y columnas, y .loc para seleccionar las 5 que finalmente utilicé: 'G', 'H', 'BA', 'SB', 'OBP', para las estadísticas ofensivas y 'Fld%', 'A', 'DP', 'E' para la defensa.

Dificultades

Debo decir que, al ser un tema que me gusta, propiamente no tuve dificultades en la elaboración o al momento de interpretar, como si me habría pasado en un tema que desconozca.

Tuve la suerte que la web de https://www.baseball-reference.com/ tiene la información bastante completa y no tuve que rellenar datos faltantes, por ejemplo.

Hablaría más quizás de limitaciones por inexperiencia, por ejemplo:

- Cometí un error importante, y es que no pude hacer gráficos porque descargué los datos en STR, cuando debí hacerlo en FLOAT.

- Por ello tampoco no saqué la moda o la mediana en las estadísticas seleccionadas, aunque solo la media fue útil y si la empleé para calcular promedio de bateo, de fildeo. El resto fueron sumatorias. Estas operaciones las hice en Excel, pero habría podido hacerlas sin problema en Pyhton.

- Sé que habría sabido cambiar las STR por FLOAT, pero me decanté por copiarlo a Excel a efectos de la presentación, sobre todo por tiempo, ya que, cuando me puse a tocar el código, se me empezaba a mover el dataframe (Se me eliminaban más filas de las deseadas, por ejemplo).

Cosas que me hubiesen gustado hacer

- Una comparativa más detallada: Un jugador jugó 24 temporadas y otro 19, esto les perjudica en ciertos apartados y beneficia en otros; Vizquel (24) sale beneficiado en las estadísticas acumulativas (sumatoria de hits o doble plays), sin embargo, disminuyendo sus temporadas a 19, su average y su promedio de fildeo habrían aumentado, ya que estas estadísticas disminuyen al envejecer los jugadores.

- Habría sido interesante hacer un análisis de Vizquel hasta su temporada 19, y de Ozzie proyectando 5 temporadas más para llegar a 24 y comparar: Quién habría sido superior en 24 temporadas y en 19.

- Esto sería más difícil, pero también sería interesante determinar que jugador tuvo mayores rivales.

Conclusiones

El análisis busqué simplificarlo, porque, para lo deseado, no vi necesario implantar técnicas más complejas. No se trataba de determinar quién era mejor, sólo si Vizquel tendría argumentos para el Salón de la Fama. Si bien es cierto que, a efectos de demostración de conocimiento, habría estado bien complementar con estadísticas descriptivas usando código y análisis de tendencias.

El análisis efectuado permitió concluir de manera efectiva.