11/24/25

Sports Analytics Semester Project

Determining if Implementing a Salary Cap in MLB would affect Competitive Advantage

Pipeline Overview & Target Grain: Rows Columns are labeled in the attached Excel Workbook. All data in the "DataDictionary" tab is pulled from the other Data tabs within the workbook, which was pulled from the following widely-used MLB data sources:

- https://www.spotrac.com/mlb/payroll/_/year/2011
- https://www.statmuse.com/mlb/ask/mlb-teams-most-playoff-wins-2010-to-2025
- https://www.mlbtraderumors.com/2024/12/largest-contract-in-franchise-history-for-each-mlb-team-2.html

ID & Mapping Strategy: Name collisions were solved by using the acronyms for each team that are widely-used across MLB (New York Yankees → NYY). For teams that have changed locations / names, used the first location first and the current location second (FLA/MIA, ATH/OAK). ANy mappings, as mentioned above, are stored within the Excel workbook.

Standardization Rules: Team names (Yankees → NYY), dates (MM/DD/YY)

Reshaping & Integration: Data is currently filtered alphabetically by team, but in the analysis, this will change based on what we are trying to see. Dropped rows: Payroll data was given per year, not average. Was able to take the average for each team and use that for simplicity reasons.

Feature & Transformation Spec: Calculated Average Annual Value for Contracts based on total value divided by years of the contract. This will be important to compare to overall payroll numbers, as the AAV contributes to a team's annual payroll, not the total contract value.

Validation Gates (QA): 30 rows of data, no duplications, ranges are realistic, fixed a variety of units (contract value, team names, dates, etc.)