11/17/25

Sports Analytics Semester Project

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

1. **Assess and summarize data sources** - clearly describe origin and contents of dataset(s).
    a. For this assignment, I will be using data from four main sources: BaseballReference, Fangraphs, Spotrac, and StatMuse. All four of these sources are widely used among the baseball analytics community as reliable sources for statistics. From BaseballReference, I will be collecting data on Win % of every team. Similarly, I will cross reference this data with that found on Fangraphs. From Spotrac, I will be collecting data regarding team payroll. Finally, from StatMuse, I will be collecting data on playoff appearances / games player per team.
2. **Evaluate data completeness and consistency** - check for missing values, coverage gaps, and inconsistent formatting or units.
    a. In my initial analysis, I mentioned data on payroll and playoff appearances would be available dating back to 1997. However, I have had trouble locating this data from a reliable source, with the most recent data dating back to 2011. As a response to this, I will be using all data dating back to only 2011 in order to stay consistent. For formatting, there may be a challenge, as payroll will be in millions, win % will be in percentage points, and playoff appearances will be in tens. As these data sets are measuring different things, I do not foresee there being any issues with these formatting inconsistencies.
3. **Identify biases or limitations** - recognize any data quality issues, anomalies, or sample biases that could affect your analysis.
    a. To start, sample biases may exist. With data from 2011 - 2025, there may be teams who have been unlucky with spending rather than just not spending at all (e.g., Mets in 2025). This could cause limitations in the data, as it may not correctly reflect that spending correlates with winning in MLB. This issue is also true for teams who may have overperformed their spending. As I am using reliable sources used widely throughout the baseball industry, I don't see there being any data quality issues. There may be anomalies, as mentioned, as teams may underperform or overperform regardless of their spending.
4. **Develop an initial cleaning plan** - outline how you will address data issues.

a. In order to keep data clean and concise, I will be importing all data into excel. By doing this, it keeps the data in one place and keeps it in a neat format. I can also analyze the data more easily, making it more likely to notice any errors or anomalies in the data.