6.1: Sourcing Open Data

Prepared by: Rob Rowland

Data Summary

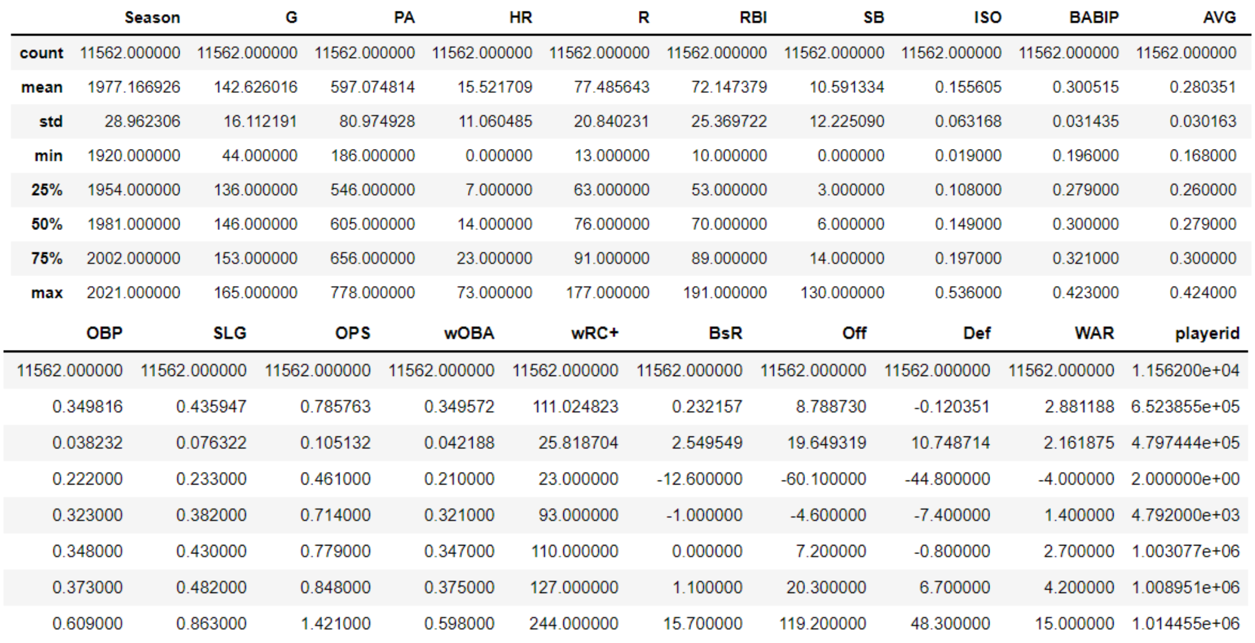
- **Data Source:**
    - This data was sourced from FanGraphs, which is a baseball statistics and analytics website that is an official partner of Major League Baseball.
- **Data Collection Method:**
    - Major League Baseball statistics are provided to FanGraphs directly from Major League Baseball. Major League Baseball employs official game scorers and statisticians to accumulate the game statistics in real time. Major League Baseball then updates their statistical database in real time.
- **Data Contents:**
    - This data set includes offensive statistics for each individual, qualified batter, by year, in the Live Ball Era of Major League Baseball (1920-present day). A qualified batter is a batter than averaged 3.1 plate appearances or more per team game over the course of a full season.
- **Data Limitations:**
    - Some statistics, such as xwOBA, where not tracked for the full durations of the data time periods. Additionally, some players that were traded during the season do not have a listed team, city, or state since they played for multiple teams within that same year.
- **Data Relevance:**
    - This data appears to meet all the qualifications for this achievement.

I chose this data because I'm passionate about sports, and baseball in particular. My interest in the analytics and statistics of sports is what helped my realize that I'd like to work in analytics as a career. The aspects of my current and former jobs that I enjoyed the most were my analytic tasks, and I've found that I regularly use my free time looking at football and baseball statistics and analytics.

Data Profile:

- Original data contains 11,562 rows and 28 columns.
- Columns:
    - Season: Year in which season was played and data recorded
    - Name: Name of player
    - Abrv: Team name abbreviated
    - City: City where player's team is based
    - State: State where player's team is based
    - Team: Player's team name
    - G: Games played
    - PA: Total plate appearances
    - HR: Total home runs
    - R: Total runs
    - RBI: Total runs batted in
    - SB: Total stolen bases

- BB%: Walk percentage of plate appearances
    - K%: Strikeout percentage of plate appearances
    - ISO: Isolated power
    - BABIP: Batting average on balls put in play
    - AVG: Batting average (Total hits / At bats)
    - OBP: On base percentage
    - SLG: Slugging percentage
    - OPS: On base plus slugging percentage
    - wOBA: Weighted on base average
    - wRC+: Weighted runs created plus
    - BsR: Base runs
    - Off: Offensive (runs + base runs)
    - Def: Defense (fielding runs + positional adjustment)
    - WAR: Wins Above Replacement
    - playerid: Number assigned to individual player
- Consistency Checks:
    - No mixed data types
    - Column xwOBA was all NaN values
        - Deleted column xwOBA
        - No other missing values
- Descriptive Statistics: Final Data Set has 11,562 rows and 27 columns



	Season	G	PA	HR	R	RBI	SB	ISO	BABIP	AVG
count	11562.000000	11562.000000	11562.000000	11562.000000	11562.000000	11562.000000	11562.000000	11562.000000	11562.000000	11562.000000
mean	1977.166926	142.626016	597.074814	15.521709	77.485643	72.147379	10.591334	0.155605	0.300515	0.280351
std	28.962306	16.112191	80.974928	11.060485	20.840231	25.369722	12.225090	0.063168	0.031435	0.030163
min	1920.000000	44.000000	186.000000	0.000000	13.000000	10.000000	0.000000	0.019000	0.196000	0.168000
25%	1954.000000	136.000000	546.000000	7.000000	63.000000	53.000000	3.000000	0.108000	0.279000	0.260000
50%	1981.000000	146.000000	605.000000	14.000000	76.000000	70.000000	6.000000	0.149000	0.300000	0.279000
75%	2002.000000	153.000000	656.000000	23.000000	91.000000	89.000000	14.000000	0.197000	0.321000	0.300000
max	2021.000000	165.000000	778.000000	73.000000	177.000000	191.000000	130.000000	0.536000	0.423000	0.424000

OBP	SLG	OPS	wOBA	wRC+	BsR	Off	Def	WAR	playerid
11562.000000	11562.000000	11562.000000	11562.000000	11562.000000	11562.000000	11562.000000	11562.000000	11562.000000	1.156200e+04
0.349816	0.435947	0.785763	0.349572	111.024823	0.232157	8.788730	-0.120351	2.881188	6.523855e+05
0.038232	0.076322	0.105132	0.042188	25.818704	2.549549	19.649319	10.748714	2.161875	4.797444e+05
0.222000	0.233000	0.461000	0.210000	23.000000	-12.600000	-60.100000	-44.800000	-4.000000	2.000000e+00
0.323000	0.382000	0.714000	0.321000	93.000000	-1.000000	-4.600000	-7.400000	1.400000	4.792000e+03
0.348000	0.430000	0.779000	0.347000	110.000000	0.000000	7.200000	-0.800000	2.700000	1.003077e+06
0.373000	0.482000	0.848000	0.375000	127.000000	1.100000	20.300000	6.700000	4.200000	1.008951e+06
0.609000	0.863000	1.421000	0.598000	244.000000	15.700000	119.200000	48.300000	15.000000	1.014455e+06

- Limitations and Ethics
    - As mentioned above, any geographical analysis will likely be limited to players who played the entire season with one team.

    - Ethically, there are no concerns in using this data. It was provided by MLB, who has agreements in place with the MLB Players Union to use all players likeness. The only personal information included in this data is the player's name.

Questions to Explore:

- Have modern medicine, techniques, and technology increased performance throughout Major League Baseball in recent years?
- Which performance statistics are affected by a player's home-team geographical location?
- How does an individual player's success affect team success? Have the most successful, single season performances come from players who also played on successful teams?
- What statistical trends from the past 100 seasons can help in predicting potential upcoming trends in Major League Baseball?