Codebook

Christopher Trzaska

2024-02-28

Overview of Data

The original raw dataset was compiled by Sean Lahman, an author and journalist with a passion for baseball. Mr. Lahman has co-authored editions of 5 sports encyclopedias in addition to writing several books on the use of statistical analysis in sports. Mr. Lahman compiled this dataset using open source research methods to provide a comprehensive look at the history of Major League Baseball and how the game has changed since its modern inception in 1876. The data set provides a list of important baseball statistics for every team's season since 1876. Baseball is notorious for its emphasis (and some might say overemphasis) on statistics. Due to the slow and deliberate pace of play and low scoring, very small changes in a team's offensive or defensive statistics can mean the difference between a World Series trip and an early elimination. This dataset initially covered a wide array of basic offensive, defensive, and season performance statistics (total runs scored, number of home runs hit, number of errors committed, etc). However it did not account for some of the newer statistics that dominate the modern game. For example, Mr. Lahman was remarkable in his ability to determine the number of hits, runs, strikeouts, and stolen bases for each team. However, more advanced statistics like batting average, on base percentage, slugging percentage, and OPS were omitted. In wrangling this data, I wanted to produce a dataset that focused primarily on teams' offensive statistics and use the basic data provided to create the more advanced statistics that are relevant to today's iteration of the sport of baseball. Doing so will provide the opportunity to analyze how individual teams improved over time as well as how the quality of MLB offenses as a whole have changed in the past 150 years.

Sources and Methodology

Sean Lahman curated the original dataset by conducting an open source research effort to compile a the history of baseball statistics. This dataset is part of a larger effort started in 1995 by Mr. Lahman to curate and disseminate raw baseball data called "the Baseball Archive". This larger effort seeks to gather data on every single MLB player who ever played a sanctioned professional baseball games in the major leagues. To support this project, Mr. Lahman looked to Major League Baseball's hall of fame in Cooperstown, NY, which provides significant archival records of historical MLB teams in addition to information on more recent years. Additionally, public websites such as baseballreference.com (which Mr. Lahman's first online baseball encyclopedia helped inspre) and mlb.com provide in depth information regarding individual and team statistics. Mr. Lahman utilized a combination of these open source data sources in the creation of this set. I have included the link to the original dataset here: https://www.openintro.org/data/index.php?data=mlb_teams

Note on Missing Values

For some variabes, Mr. Lahaman was unable to find data for that time frame. In some cases, this is because recordkeeping in the early days of baseball for stats beyond basic run, hit, and out totals was limited. In other cases, the structure of the league changed which altered how seasons were conducted. For example, initially Major League Baseball consisted of just 8 teams playing a series of sanctioned games with no final championship. As the league expanded and new teams were created, the league was divided into sub-leagues and divisions that offered the chance for teams to win a smaller title to advance to the playoffs and compete for the "World Championship". Thus, there is no information about who won a given division before that division was created. All values in the dataset where there is no availible information are left blank and marked with NA.

Itemized Presentation of Variables

Variable name : ***year***

Variable type : character

Description: Year of play of team being observed.

Variable name : ***team_name***

Variable type : character

Description: Name of team being observed.

Variable name : ***ball_park***

Variable type : character

Description: Name of home stadium where team observed played that year.

Variable name : ***league_id***

Variable type : factor

Description: Name of sub-league the team observed competed in. Options are either NL (National League) or AL (American League)

Var1	Freq
AL	1280
NL	1504

Variable name : ***division_id***

Variable type : factor

Description: Name of sub-division of the sub-league the team observed competed in. Options are either E (East), C (Central), W (West). Divisions were created in 1969.

Var1	Freq
C	285
E	588
W	565

Variable name : ***games_played***

Variable type : numerical

Description: Total number of games the observed team played during that year's season.

mean	median	min	max
153.1185	161	57	165

Variable name : ***wins***

Variable type : numeric

Description: How many total games the observed team won during the regular season.

mean	median	min	max
153.1185	161	57	165

Variable name : ***winning_per***

Variable type : numeric

Description: What percentage of total regular season games the observed team won.

mean	median	min	max
0.4973729	0.5	0.1298701	0.7878788

Variable name : ***wild_card_winner***

Variable type : factor

Description: Was the team able to secure a playoff spot via the wild card. Starting in 2012, the two highest seeded teams that did not win their respective divisions were granted a playoff berth via the wild card as a way to expand the playoff format. That number was later increased to the top 3 teams starting in 2022. Options are either Y (Yes) or N (No)

Var1	Freq
N	698
Y	76

Variable name : ***division_winner***

Variable type : factor

Description: Whether the observed team won their respective division. MLB was reorganized into the divisional framework in 1969. Each MLB team was placed into either the American League or National League and then assigned a geographical-based division. The team with the best regular season record at the end of the season wins their division and is granted a playoff berth. Options are Y (Yes) team won division or N (No) they didnt.

Var1	Freq
N	1150
Y	260

Variable name : ***league_winner***

Variable type : factor

Description: Whether the observed team won their league. MLB teams are divided into 2 leagues, American and National. The winner of the league is determined by which team wins the League Championship Series. The winner of each league has advanced to the "World Series" since 1903 to determine the MLB champion for that year. Options are either Y (Yes) team won their respective league or N (No) they did not.

Var1	Freq
N	2493
Y	263

Variable name : ***world_series_winner***

Variable type : factor

Description: Whether the observed team won the World Series for that year. Since 1903, the winners of the American and National leagues have competed for the World Series title to be crowned champions of Major League Baseball. Options are Y (Yes) team won the World Series that year or N (No) they did not.

Var1	Freq
N	2416
Y	120

Variable name : ***runs_scored***

Variable type : numeric

Description: How many total runs the observed team scored during the regular season.

mean	median	min	max
687.5765	693	219	1220

Variable name : ***at_bats***

Variable type : numeric

Description: How many total times an offensive player stepped into the batter's box to face a pitch during the regular season.

mean	median	min	max
5227.865	5422	1752	5781

Variable name : ***hits***

Variable type : numeric

Description: How many total hits the observed team achieved during the regular season. A hit is any time a batter safely reaches or passes first base after hitting the ball into fair territory with neither the benefit of an error nor a fielder's choice.

mean	median	min	max
1367.211	1398	390	1783

Variable name : ***singles***

Variable type : numeric

Description: How many total singles the observed team achieved during the regular season. A single is any time a batter safely reache first base after hitting the ball into fair territory with neither the benefit of an error nor a fielder's choice.

mean	median	min	max
976.8272	992	221	1345

Variable name : ***doubles***

Variable type : numeric

Description: How many total doubles the observed team achieved during the regular season. A single is any time a batter safely reaches second base after hitting the ball into fair territory with neither the benefit of an error nor a fielder's choice.

mean	median	min	max
234.6541	238	39	376

Variable name : ***triples***

Variable type : numeric

Description: How many total triples the observed team achieved during the regular season. A single is any time a batter safely reaches triple base after hitting the ball into fair territory with neither the benefit of an error nor a fielder's choice.

mean	median	min	max
45.47198	40	3	150

Variable name : ***homeruns***

Variable type : numeric

Description: How many total home runs the observed team achieved during the regular season. A single is any time a batter safely rounds all 3 bases and touches home plate after hitting the ball into fair territory with neither the benefit of an error nor a fielder's choice.

mean	median	min	max
110.2575	114	0	307

Variable name : ***walks***

Variable type : numeric

Description: How many total walks the observed team achieved during the regular season. A walk is any time a pitcher throws a total of 4 balls out of the strike zone to a batter during an at bat that the batter does not swing at, make contact with, or put in play.

mean	median	min	max
487.389	499	18	835

Variable name : ***strikeouts_by_batters***

Variable type : numeric

Description: How many total times did the observed team's batters strike out during the regular season. A strike out is any time a batter records 3 strikes in an at bat. A strike is a pitch thrown over the plate and in between the batter's knees and mid chest, a pitch hit into foul territory, or swung at and missed. A batter cannot strike out on a foul ball unless it is caught by the catcher.

mean	median	min	max
781.8468	800	35	1595

Variable name : ***stolen_bases***

Variable type : numeric

Description: How many total times did the observed team's runners advance to the next base without being batted over by a ball in play and without the assistance of a throwing or fielding error by the defense.

mean	median	min	max
106.2478	92	13	441

Variable name : ***caught_stealing***

Variable type : numeric

Description: How many total times did the observed team's runners get thrown out by the defense while attempting to advance to the next base without being batted over by a ball in play and without the assistance of a throwing or fielding error by the defense.

mean	median	min	max
47.54624	44	3	191

Variable name : ***batters_hit_by_pitch***

Variable type : numeric

Description: How many total times did the observed team's batters get hit by a pitch. If a batter is standing in the batter's box and is hit by a pitched ball, he is awarded first base.

mean	median	min	max
45.17055	43	7	160

Variable name : ***sacrifice_flies***

Variable type : numeric

Description: How many total times did the observed team complete a sacrifice fly. A sacrifice fly occurs when a batter hits a fly ball that is caught for an out but allows runners to advance to the next base. Runners are allowed to advance if they start their sprint to the next base at the moment the batted ball is caught in the outfield.

mean	median	min	max
44.23338	44	7	77

Variable name : ***on_base_per***

Variable type : numeric

Description: The percentage of at bats resulted in the observed team's batters reaching base safely in any capacity (hit (- home runs), walk, hit by pitch).

mean	median	min	max
0.3591175	0.3596261	0.2348624	0.4711398

Variable name : ***avg_obp***

Variable type : numeric

Description: The average on base percentage of every team across the entire MLB for a given year.

mean	median	min	max
0.3591175	0.3609605	0.2746571	0.4210828

Variable name : ***obp_status***

Variable type : factor

Description: Identifies whether the team's total on base percentage was "Over" or "Under" the MLB's total average on base percentage for that year.

Var1	Freq
Over	1350
Under	1434

Variable name : ***slugging_perc***

Variable type : numeric

Description: Slugging percentage is a measure of a player's hitting productivity. It is calculated by dividing the total bases the player accumulated via hitting by their total at bats. Each type of hit is awarded a multiplier according to their strength. Singles get a 1x multiplier, doubles a 2x, triples a 3x, and home runs a 4x. This is the sum of the total bases of offense the player is resonsible for producing. This sum is then divided by the total number at bats they had and the result is their slugging percentage.

mean	median	min	max
0.3857155	0.3879206	0.2607484	0.495457

Variable name : ***avg_slugging***

Variable type : numeric

Description: The average slugging percentage of every team across the entire MLB for a given year.

mean	median	min	max
0.3857155	0.3904014	0.3048238	0.4370553

Variable name : ***obp_status***

Variable type : factor

Description: Identifies whether the team's total slugging percentage was "Over" or "Under" the MLB's total average slugging percentage for that year.

Var1	Freq
Over	1367
Under	1417

Variable name : ***ops***

Variable type : numeric

Description: OPS stands for "on base plus slugging" and it calculated by adding together a batter's slugging and on base percentages. It provides a more holistic look at the offensive production of a player or team by including their ability to get on base in ways beyond getting a hit. A player's total bases via hitting are captures in the slugging percentage, their ability to work walks and get on base in unconventional ways is captured in their on base percentage. Thus this advanced statistic can provide a much deeper understanding of offensive production.

mean	median	min	max
0.744833	0.7493409	0.4958716	0.9539487

Variable name : ***avg_ops***

Variable type : numeric

Description: The average "on base plus slugging" score of every team across the entire MLB for a given year.

mean	median	min	max
0.744833	0.7496993	0.5939741	0.8553123

Variable name : ***ops_status***

Variable type : factor

Description: Identifies whether the team's total "on base percentage was"on base plus slugging" score was "Over" or "Under" the MLB's total average OPS for that year.

Var1	Freq
Over	1371
Under	1413

Variable name : ***batting_avg***

Variable type : numeric

Description: Represents how many hits a player or team achieved divided by their total number of at bats. This is a quintessential baseball offense statistic that represents how often a batter records a hit in their at bats.

mean	median	min	max
0.2612924	0.2603309	0.2076551	0.3498428

Variable name : ***avg_ba***

Variable type : numeric

Description: The average of all observed team's batting averages in a given year.

mean	median	min	max
0.2612924	0.259777	0.2364674	0.3086536

Variable name : ***ba_status***

Variable type : factor

Description: Identifies whether the teams' batting average was "Over" or "Under" the MLB's total average batting average for that year.

Var1	Freq
Over	1379
Under	1405