APPENDIX A – LAHMAN DATASET

Table Name/ Description	*AllstarFull* Description: All Star appearances by players A data frame with 5375 observations on the following 8 variables			
Variable Name	**Description**	**Statistical Summary/Type**	**Errors**	**Used in final dataset?**
playerID	Player ID code	Unique Values: 1867 Type: Object	None	Yes/No
yearID	Year	Min: 1933 Max: 2019 Unique Values: 87 Type: float64	Missing Values: 1	
gameNum	Game number (for years in which more than one game was played)	Min: 0 Max: 2 Unique Values: 3 Type: float64	Missing Values: 1	
gameID	Game ID code	Unique Values: 90 Type: Object	Missing Values: 50	
teamID	Team; a factor	Unique Values: 44 Type: Object	None	
lgID	League; a factor with levels AL, NL	Unique Values: 2 Type: Object	None	
GP	Game played (zero if player did not appear in game)	Min: 0 Max: 1 Unique Values: 2 Type: int64	None	
startingPos	If the player started, what position he played	Min: 1 Max: 10 Unique Values: 10 Type: float64	Missing Values: 3684	
Table Name/ Description	*Appearances* Description: Data on player appearances A data frame with 107357 observations on the following 21 variables			
Variable Name	**Description**	**Statistical Summary/Type**	**Errors**	**Used in final dataset?**
yearID	Year	Min: 1871 Max: 2019 Unique Values: 149 Type: int64	None	
teamID	Team; a factor	Unique Values: 149 Type: Object	None	
lgID	League; a factor with levels AA, AL, FL, NL, PL, UA	Unique Values: 6 Type: Object	Missing Values: 737	
playerID	Player ID code	Unique Values: 19690 Type: Object	None	

gameNum	Game number (for years in which more than one game was played)	Min: 0 Max: 2 Unique Values: 3 Type: float64	Missing Values: 1	
gameID	Game ID code	Unique Values: 90 Type: Object	Missing Values: 50	
G_all	Total games played	Min: 1 Max: 165 Unique Values: 165 Type: int64	None	
GS	Games started	Min: 0 Max: 164 Unique Values: 165 Type: int64	Missing Values: 8962	
G_batting	Games in which player batted	Min: 0 Max: 165 Unique Values: 166 Type: int64	None	
G_defense	Games in which player appeared on defense	Min: 0 Max: 165 Unique Values: 166 Type: int64	Missing Values: 7698	
G_p	Games as pitcher	Min-Max: 0 – 106 Unique Values: 95	None	
G_c	Games as catcher	Min-Max: 0 – 160 Unique Values: 157	None	
G_1b G_2b G_3b	Games as first baseman Games as second baseman Games as third baseman	Min-Max: 0 – 162 Min-Max: 0 – 163 Min-Max: 0 – 164	None	
G_ss	Games as shortstop	Min-Max: 0 – 165	None	
G_lf G_cf G_rf G_of	Games as leftfielder Games as centerfielder Games as right fielder Games as outfielder	Min-Max: 0 – 163 Min-Max: 0 – 162 Min-Max: 0 – 162 Min-Max: 0 – 164	None	
G_dh	Games as designated hitter	Min-Max: 0 – 162 Unique Values: 157	Missing Values: 1264	
G_ph	Games as pinch hitter	Min-Max: 0 – 109 Unique Values: 94	Missing Values: 8962	
G_pr	Games as pinch runner	Min-Max: 0 – 92 Unique Values: 50	Missing Values: 8962	
Table Name/ Description	*AwardsManagers* Description: Award information for managers awards A data frame with 179 observations on the following 6 variables			
Variable Name	**Description**	**Statistical Summary/Type**	**Errors**	**Used in final dataset?**
playerID	Manager (player) ID code	Unique Values: 83 Type: Object	None	Yes/No
awardID	Name of award won	Unique Values: 2 Type: Object	None	
yearID	Year	Min: 1936	None	

		Max: 2016 Unique Values: 81 Type: int64		
lgID	League; a factor with levels AL, NL	Unique Values: 2 Type: Object	None	
gameNum	Game number (for years in which more than one game was played)	Min: 0 Max: 2 Unique Values: 3 Type: float64	Missing Values: 1	
gameID	Game ID code	Unique Values: 90 Type: Object	Missing Values: 50	
teamID	Team; a factor	Unique Values: 44 Type: Object	None	
tie	Award was a tie (Y or N)	Unique Values: 1 Type: Object	Missing Values: 177 *Null values should be converted to N or zeros*	
notes	Notes about the award	Unique Values: 1 Type: Object	Missing Values: 178	
Table Name/ Description	*AwardsPlayers* Description: Award information for players awards A data frame with 6236 observations on the following 6 variables			
Variable Name	**Description**	**Statistical Summary/Type**	**Errors**	**Used in final dataset?**
playerID	Player ID code	Unique Values: 1360 Type: Object	None	Yes/No
awardID	Name of award won	Unique Values: 29 Type: Object	None	
yearID	Year	Min: 1936 Max: 2016 Unique Values: 81 Type: int64	None	
lgID	League; a factor with levels AL, NL, ML, NL	Unique Values: 4 Type: Object	None	
tie	Award was a tie (Y or N)	Unique Values: 1 Type: Object	Missing Values: 6187 *Null values should be converted to N or zeros*	
notes	Notes about the award	Unique Values: 29 Type: Object	Missing Values: 1468	
Table Name/ Description	*AwardsShareManagers* Description: Award voting for managers awards A data frame with 425 observations on the following 7 variables			
Variable Name	**Description**	**Statistical Summary/Type**	**Errors**	**Used in final dataset?**

awardID	name of award votes were received for	Unique Values: 1 Type: Object	None	Yes/No
yearID	Year	Min: 1983 Max: 2016 Unique Values: 34 Type: float64	None	
lgID	League; a factor with levels AL, NL	Unique Values: 2 Type: Object	None	
playerID	Manager (player) ID code	Unique Values: 112 Type: Object	None	
pointsWon	Number of points received	Min: 1 Max: 154 Unique Values: 118 Type: int64	None	
pointsMax	Maximum number of points possible	Min: 24 Max: 160 Unique Values: 6 Type: float64	Missing Values: 11	
votesFirst	Number of first place votes	Min: 0 Max: 30 Unique Values: 30 Type: int64	None	
Table Name/ Description	*AwardsSharePlayers* Description: Award voting for managers awards A data frame with 6879 observations on the following 7 variables			
Variable Name	**Description**	**Statistical Summary/Type**	**Errors**	**Used in final dataset?**
awardID	name of award votes were received for	Unique Values: 3 Type: Object	None	Yes/No
yearID	Year	Min: 1911 Max: 2016 Unique Values: 98 Type: int64	None	
lgID	League; a factor with levels AL ML NL	Unique Values: 3 Type: Object	None	
playerID	Player ID code	Unique Values: 2441 Type: Object	None	
pointsWon	Number of points received	Min: 1 Max: 154 Unique Values: 349 Type: float64	None	
pointsMax	Maximum number of points possible	Min: 24 Max: 160 Unique Values: 21 Type: int64	None	
votesFirst	Number of first place votes	Min: 0 Max: 30 Unique Values: 39 Type: float64	Missing Values: 358	

Table Name/ Description	*Batting* Description: Batting table - batting statistics A data frame with 107429 observations on the following 22 variables			
Variable Name	**Description**	**Statistical Summary/Type**	**Errors**	**Used in final dataset?**
playerID	Player ID code	Unique Values: 19689 Type: Object	None	Yes/No
yearID	Year	Min: 1933 Max: 2019 Unique Values: 87 Type: float64	None	
stint	player's stint (order of appearances within a season)	Min: 1 Max: 5 Unique Values: 5 Type: int64	None	
teamID	Team; a factor	Unique Values: 44 Type: Object	None	
lgID	League; a factor with levels AA AL FL NL PL UA	Unique Values: 2 Type: Object	Missing Values: 738	
G	Games: number of games in which a player played	Type: int64	None	
AB	At Bats	Type: int64	None	
R	Runs	Type: int64	None	
H	Hits: times reached base because of a batted, fair ball without error by the defense	Type: int64	None	
2B	Doubles: hits on which the batter reached second base safely	Type: int64	None	
3B	Triples: hits on which the batter reached third base safely	Type: int64	None	
HR	Homeruns	Type: int64	None	
RBI	Runs Batted In	Type: float64	Missing Values: 756	
SB	Stolen Bases	Type: float64	Missing Values: 2368	
CS	Caught Stealing	Type: float64	Missing Values: 23541	
BB	Base on Balls	Type: int64	None	
SO	Strikeouts	Type: float64	Missing Values: 2100	
IBB	Intentional walks	Type: float64	Missing Values: 36651	
HBP	Hit by pitch	Type: float64	Missing Values: 2817	
SH	Sacrifice hits	Type: float64	Missing Values: 6069	
SF	Sacrifice flies	Type: float64	Missing Values: 36104	

GIDP	Grounded into double plays	Type: float64	Missing Values: 25441	
Table Name/ Description	*BattingPost* Description: Post season batting statistics A data frame with 14750 observations on the following 22 variables			
Variable Name	**Description**	**Statistical Summary/Type**	**Errors**	**Used in final dataset?**
playerID	Player ID code	Unique Values: 19689 Type: Object	None	Yes/No
yearID	Year	Min: 1933 Max: 2019 Unique Values: 87 Type: float64	None	
stint	player's stint (order of appearances within a season)			
teamID	Team; a factor	Unique Values: 44 Type: Object	None	
lgID	League; a factor with levels AA AL NL	Unique Values: 2 Type: Object	None	
G	Games: number of games in which a player played	Type: int64	None	
AB	At Bats	Type: int64	None	
R	Runs	Type: int64	None	
H	Hits: times reached base because of a batted, fair ball without error by the defense	Type: int64	None	
2B	Doubles: hits on which the batter reached second base safely	Type: int64	None	
3B	Triples: hits on which the batter reached third base safely	Type: int64	None	
HR	Homeruns	Type: int64	None	
RBI	Runs Batted In	Type: float64	None	
SB	Stolen Bases	Type: float64	None	
CS	Caught Stealing	Type: float64	Missing Values: 201	
BB	Base on Balls	Type: int64	None	
SO	Strikeouts	Type: float64	None	
IBB	Intentional walks	Type: float64	None	
HBP	Hit by pitch	Type: float64	Missing Values: 201	
SH	Sacrifice hits	Type: float64	Missing Values: 201	
SF	Sacrifice flies	Type: float64	Missing Values: 201	
GIDP	Grounded into double plays	Type: float64	Missing Values: 201	

Table Name/ Description	*CollegePlaying* Description: Information on schools' players attended, by player A data frame with 17350 observations on the following 3 variables			
Variable Name	**Description**	**Statistical Summary/Type**	**Errors**	**Used in final dataset?**
playerID	Player ID code	Unique Values: 6575 Type: Object	None	Yes/No
schoolID	school ID code	Unique Values: 1038 Type: Object	None	
yearID	Year player attended school	Min: 1864 Max: 2014 Unique Values: 151 Type: int64	None	
Table Name/ Description	*Fielding* Description: Fielding table A data frame with 143046 observations on the following 18 variables			
Variable Name	**Description**	**Statistical Summary/Type**	**Errors**	**Used in final dataset?**
playerID	Player ID code	Unique Values: 19491 Type: Object	None	Yes/No
yearID	Year	Min: 1871 Max: 2019 Unique Values: 149 Type: int64	None	
stint	player's stint (order of appearances within a season)	Min: 1 Max: 5 Unique Values: 5 Type: int64	None	
teamID	Team; a factor	Unique Values: 149 Type: Object	None	
lgID	League; a factor with levels AA AL FL NL PL UA	Unique Values: 6 Type: Object	Missing Values: 1513	
G	Games	Type: int64	None	
GS	Games Started	Type: float64	Missing Values: 46157	
InnOuts	Time played in the field expressed as outs	Type: float64	Missing Values: 29929	
PO	Putouts	Type: int64	None	
A	Assists	Type: int64	None	
E	Errors	Type: float64	Missing Values: 1	
DP	Double Plays	Type: int64	None	
PB	Passed Balls (by catchers)	Type: float64	Missing Values: 131443	
WP	Wild Pitches (by catchers)	Type: float64	Missing Values: 141877	
SB	Opponent Stolen Bases (by catchers)	Type: float64	Missing Values: 134230	

CS	Opponents Caught Stealing (by catchers)	Type: float64	Missing Values: 134230	
ZR	Zone Rating	Type: float64	Missing Values: 141877	
Table Name/ Description	*FieldingOF* DescriptionOutfield position data: information about positions played in the outfield A data frame with 33279 observations on the following 18 variables			
Variable Name	**Description**	**Statistical Summary/Type**	**Errors**	**Used in final dataset?**
playerID	Player ID code	Unique Values: 3513 Type: Object	None	Yes/No
yearID	Year	Min: 1933 Max: 2019 Unique Values: 85 Type: int64	None	
stint	player's stint (order of appearances within a season)	Min: 1 Max: 5 Unique Values: 5 Type: int64	None	
Glf	Games played in left field	Min: 0 Max: 156 Unique Values: 157 Type: float64	Missing Values: 37	
Gcf	Games played in center field	Min: 0 Max: 162 Unique Values: 159 Type: float64	Missing Values: 37	
Grf	Games played in right field	Min: 0 Max: 160 Unique Values: 159 Type: float64	Missing Values: 43	
Table Name/ Description	*FieldingOFsplit* Description: Outfield position data: information about positions played in the outfield A data frame with 5375 observations on the following 8 variables			
Variable Name	**Description**	**Statistical Summary/Type**	**Errors**	**Used in final dataset?**
playerID	Player ID code	Unique Values: 3529 Type: Object	None	Yes/No
yearID	Year	Min: 1954 Max: 2019 Unique Values: 66 Type: int64	None	
stint	player's stint (order of appearances within a season)	Min: 1 Max: 4 Unique Values: 4 Type: int64	None	
teamID	Team; a factor	Unique Values: 149 Type: Object	None	

lgID	League; a factor with levels AA AL FL NL PL UA	Unique Values: 6 Type: Object	Missing Values: 1513	
POS	Position	Unique Values: 3 Type: Object		
G	Games	Type: int64	None	
GS	Games Started	Type: int64	None	
InnOuts	Time played in the field expressed as outs	Type: int64	None	
PO	Putouts	Type: int64	None	
A	Assists	Type: int64	None	
E	Errors	Type: int64	None	
DP	Double Plays	Type: int64	None	
PB	Passed Balls (by catchers)	Type: float64	Missing Values: 33279	
WP	Wild Pitches (by catchers)	Type: float64	Missing Values: 33279	
SB	Opponent Stolen Bases (by catchers)	Type: float64	Missing Values: 33279	
CS	Opponents Caught Stealing (by catchers)	Type: float64	Missing Values: 33279	
ZR	Zone Rating	Type: float64	Missing Values: 33279	
Table Name/ Description	*HallOfFame* Description: Hall of Fame table. This is composed of the voting results for all candidates nominated for the Baseball Hall of Fame. A data frame with 4191 observations on the following 9 variables			
Variable Name	**Description**	**Statistical Summary/Type**	**Errors**	**Used in final dataset?**
playerID	Player ID code	Unique Values: 1279 Type: Object	None	Yes/No
yearID	Year of ballot	Min: 1933 Max: 2019 Unique Values: 80 Type: float64	None	
votedBy	Method by which player was voted upon	Unique Values: 9 Type: Object	None	
ballots	Total ballots cast in that year	Unique Values: 74	Missing Values: 197	
needed	Number of votes needed for selection in that year	Unique Values: 65	Missing Values: 354	
votes	Total votes received	Unique Values: 367	Missing Values: 197	
inducted	Whether player was inducted by that vote or not (Y or N)	Unique Values: 2 Type: Object	None	
category	Category of candidate; a factor with levels Manager Pioneer/Executive Player Umpire	Unique Values: 4 Type: Object	None	
needed_note	Explanation of qualifiers for special elections	Unique Values: 2 Type: Object	Missing Values: 4034	

Table Name/ Description	*HomeGames* Description: Data mapping teams to the stadiums they played regular season games in as the home team; A data frame with 3108 observations on the following 9 variables			
Variable Name	**Description**	**Statistical Summary/Type**	**Errors**	**Used in final dataset?**
year .key	Year	Min-Max: 1871 - 2019 Unique Values: 149 Type: int64	None	Yes/No
league .key	League; a factor with levels AA AL FL NL PL UA	Unique Values: 6 Type: object	Missing Values: 77	
team .key	Team; a factor	Unique Values: 148 Type: object	None	
park .key	Unique identifier for each ballpark	Unique Values: 254 Type: object	None	
span .first	First date the park began acting as home field for the team	Unique Values: 1099 Type: object	Inconsistencies in time/date formats	
span .last	Last date the park began acting as home field for the team	Unique Values: 1137 Type: object	Inconsistencies in time/date formats	
games	Total games in this time span	Min-Max: 1 - 89 Unique Values: 87 Type: int64	None	
openings	openings Total opening in this time span	Min-Max: 0 - 83 Unique Values: 84 Type: int64	None	
attendance	Total attendance in this time span	Unique Values: 2665 Type: int64	None	
Table Name/ Description	*Managers* Description: Managers table: information about individual team managers, teams they managed and some basic statistics for those teams in each year A data frame with 3536 observations on the following 10 variables			
Variable Name	**Description**	**Statistical Summary/Type**	**Errors**	**Used in final dataset?**
playerID	Manager ID code	Unique Values: 714 Type: Object	None	Yes/No
yearID	Year	Min: 1871 Max: 2019 Unique Values: 149 Type: float64	Missing Values: 1	
teamID	Team; a factor	Unique Values: 44 Type: Object	None	
lgID	League; a factor with levels AA AL FL NL PL UA	Unique Values: 6 Type: Object	Missing Values: 67	
inseason	Managerial order. Zero if the individual managed the team the entire year. Otherwise denotes where the manager appeared in the managerial	Unique Values: 9 Type: int64	None	

	order (1 for first manager, 2 for second, etc.)			
G	Games Managed	Min-Max: 1 – 165 Unique Values: 165 Type: int64	None	
W	Wins	Min-Max: 1 – 116 Unique Values: 114 Type: int64	None	
L	Loses	Min-Max: 1 – 120 Unique Values: 119 Type: int64	None	
rank	Team's final position in standings that year	Unique Values: 12 Type: float64	None	
plyMgr	Player Manager (denoted by 'Y'); a factor with levels N Y	Unique Values: 2 Type: object	None	
Table Name/ Description	*ManagersHalf* Description: Split season data for managers A data frame with 93 observations on the following 10 variables			
Variable Name	**Description**	**Statistical Summary/Type**	**Errors**	**Used in final dataset?**
playerID	Manager ID code	Unique Values: 54 Type: Object	None	Yes/No
yearID	Year	Only Two Values : 1892 - 1981 Unique Values: 2 Type: int64	None	
teamID	Team; a factor	Unique Values: 33 Type: Object	None	
lgID	League; a factor with levels AL NL	Unique Values: 2 Type: Object	None	
inseason	Managerial order. Zero if the individual managed the team the entire year. Otherwise denotes where the manager appeared in the managerial order (1 for first manager, 2 for second, etc.)	Unique Values: 5 Type: int64	None	
half	First or second half of season	Unique Values: 2 Type: int64	None	
G	Games Managed	Min-Max: 1 – 165 Unique Values: 40 Type: int64	None	
W	Wins	Min-Max: 1 – 116 Unique Values: 42 Type: int64	None	
L	Loses	Min-Max: 1 – 120 Unique Values: 34 Type: int64	None	
rank	Team's final position in standings that year	Unique Values: 12 Type: int64	None	

Table Name/ Description	*Parks* Description: Name and location data for baseball stadiums A data frame with 255 observations on the following 6 variables			
Variable Name	**Description**	**Statistical Summary/Type**	**Errors**	**Used in final dataset?**
park .key	unique identifier for each ballpark	Unique Values: 255 Type: Object	None	Yes/No
park .name	the name of the ballpark	Unique Values: 243 Type: Object	None	
park .alias	a semicolon delimited list of other names for the ballpark if they exist	Unique Values: 57 Type: Object	Missing Values: 196	
city	city where the ballpark is located	Unique Values: 87 Type: Object	None	
state	state where the ballpark is located	Unique Values: 37 Type: Object	Missing Values: 1	
country	country where the ballpark is located	Unique Values: 7 Type: Object	None	
Table Name/ Description	*People* Description: People table - Player names, DOB, and biographical info. This file is to be used to get details about players listed in the Batting, Pitching, and other files where players are identified only by playerID A data frame with 19878 observations on the following 26 variables			
Variable Name	**Description**	**Statistical Summary/Type**	**Errors**	**Used in final dataset?**
playerID	A unique code assigned to each player. The playerID links the data in this file with records on players in the other files	Unique Values: 20087 Type: Object	None	Yes/No
birthYear birthMonth birthDay	Year, Month, and Day player was born	Type: float64 Unique Values: 169 Unique Values: 12 Unique Values: 31	Missing Values: 114 252 424	
birthCountry birthState birthCity	Country, State, and City where player was born	Type: Object Unique Values: 57 Unique Values: 296 Unique Values: 4883	Missing Values: 61 555 172	
deathYear deathMonth deathDay	Year, Month, and Day player died	Type: float64 Unique Values: 149 Unique Values: 12 Unique Values: 31	Missing Values: 10253 10254 10255	
deathCountry deathState deathCity	Country, State, and City where player died	Type: Object Unique Values: 25 Unique Values: 107 Unique Values: 2669	Missing Values: 10255 10305 10260	
nameFirst	Player's first name	Type: Object Unique Values: 2529	Missing Values: 37	

nameLast nameGiven	Player's last name Player's given name (typically first and middle)	Unique Values: 10237 Unique Values: 13337	None 37	
weight	Player's weight in pounds	Unique Values: 153	Missing Values: 817	
height	Player's height in inches	Unique Values: 23	Missing Values: 737	
bats	a factor: Player's batting hand (left (L), right (R), or both (B))	Unique Values: 3 Type: Object	Missing Values: 1180	
throws	a factor: Player's throwing hand (left(L) or right(R))	Unique Values: 3 Type: Object	Missing Values: 977	
debut	Date that player made first major league appearance	Unique Values: 10572 Type: Object	Missing Values: 195	
finalGame	Date that player made first major league appearance (blank if still active)	Unique Values: 9479 Type: Object	Missing Values: 195	
retroID	ID used by retrosheet	Unique Values: 20030 Type: Object	Missing Values: 57	
bbrefID	ID used by Baseball Reference website	Unique Values: 20084 Type: Object	Missing Values: 3	
birthDate	Player's birthdate, in as Date format	Type: date/time	Inconsistencies in date formats	
deathDate	Player's death date, in as Date format	Type: date/time	Inconsistencies in date formats	
Table Name/ Description	*Pitching* Description: Pitching table A data frame with 47628 observations on the following 30 variables			
Variable Name	**Description**	**Statistical Summary/Type**	**Errors**	**Used in final dataset?**
playerID	Player ID code	Unique Values: 9845 Type: Object	None	Yes/No
yearID	Year	Unique Values: 149 Type: int64	Missing Values: 1	
stint	player's stint (order of appearances within a season)	Unique Values: 5 Type: int64	None	
teamID	Team; a factor	Unique Values: 149 Type: Object	None	
lgID	League; a factor with levels AA AL FL NL PL UA	Unique Values: 6 Type: Object	Missing Values: 132	
W	Wins	Unique Values: 54 Type: int64	None	
L	Losses	Unique Values: 43 Type: int64	None	
G	Games	Unique Values: 94 Type: int64	None	
GS	Games Started	Unique Values: 75 Type: int64	None	
CG	Complete Games	Unique Values: 74 Type: int64	None	

SHO	Shutouts	Unique Values: 15 Type: int64	None	
SV	Saves	Unique Values: 57 Type: int64	None	
IPouts	Outs Pitched (innings pitched x 3)	Unique Values: 1311 Type: int64	None	
H	Hits	Unique Values: 534 Type: int64	None	
ER	Earned Runs	Unique Values: 216 Type: int64	None	
HR	Homeruns	Unique Values: 48 Type: int64	None	
BB	Walks	Unique Values: 209 Type: int64	None	
SO	Strikeout	Unique Values: 338 Type: int64	None	
BAOpp	Opponent's Batting Average	Unique Values: 448 Type: float64	Missing Values: 4441	
ERA	Earned Run Average	Unique Values: 1168 Type: float64	Missing Values: 94	
IBB	Intentional Walks	Unique Values: 22 Type: float64	Missing Values: 14578	
WP	Wild Pitches	Unique Values: 61 Type: int64	None	
HBP	Batters Hit By Pitch	Unique Values: 42 Type: float64	Missing Values: 734	
BK	Balks	Unique Values: 16 Type: int64	None	
BFP	Batters faced by Pitcher	Unique Values: 1736 Type: float64	None	
GF	Games Finished	Unique Values: 78 Type: int64	None	
R	Runs Allowed	Unique Values: 345 Type: int64	None	
SH	Sacrifices by opposing batters	Unique Values: 25 Type: float64	Missing Values: 19187	
SF	Sacrifice flies by opposing batters	Unique Values: 18 Type: float64	Missing Values: 19187	
GIDP	Grounded into double plays by opposing batter	Unique Values: 44 Type: float64	Missing Values: 23018	
Table Name/ Description	*Pitching Post* Description: Post season pitching statistics A data frame with 5798 observations on the following 30 variables			
Variable Name	**Description**	**Statistical Summary/Type**	**Errors**	**Used in final dataset?**
playerID	Player ID code	Unique Values: 1829 Type: Object	None	Yes/No
yearID	Year	Unique Values: 123 Type: int64	None	
round	Level of playoffs	Unique Values: 14	None	

		Type: int64		
teamID	Team; a factor	Unique Values: 48 Type: Object	None	
lgID	League; a factor with levels AA AL FL NL PL UA	Unique Values: 3 Type: Object	None	
W	Wins	Unique Values: 5 Type: int64	None	
L	Losses	Unique Values: 5 Type: int64	None	
G	Games	Unique Values: 8 Type: int64	None	
GS	Games Started	Unique Values: 8 Type: int64	None	
CG	Complete Games	Unique Values: 8 Type: int64	None	
SHO	Shutouts	Unique Values: 4 Type: int64	None	
SV	Saves	Unique Values: 5 Type: int64	None	
IPouts	Outs Pitched (innings pitched x 3)	Unique Values: 90 Type: int64	None	
H	Hits	Unique Values: 39 Type: int64	None	
ER	Earned Runs	Unique Values: 20 Type: int64	None	
HR	Homeruns	Unique Values: 6 Type: int64	None	
BB	Walks	Unique Values: 19 Type: int64	None	
SO	Strikeout	Unique Values: 31 Type: int64	None	
BAOpp	Opponent's Batting Average	Unique Values: 306 Type: float64	Missing Values: 71	
ERA	Earned Run Average	Unique Values: 334 Type: float64	Missing Values: 33	
IBB	Intentional Walks	Unique Values: 5 Type: float64	Missing Values: 50	
WP	Wild Pitches	Unique Values: 6 Type: int64	Missing Values: 50	
HBP	Batters Hit By Pitch	Unique Values: 5 Type: float64	Missing Values: 50	
BK	Balks	Unique Values: 2 Type: int64	Missing Values: 50	
BFP	Batters faced by Pitcher	Unique Values: 109 Type: float64	Missing Values: 50	
GF	Games Finished	Unique Values: 7 Type: int64	None	
R	Runs Allowed	Unique Values: 28 Type: int64	None	
SH	Sacrifices by opposing batters	Unique Values: 8 Type: float64	Missing Values: 50	

SF	Sacrifice flies by opposing batters	Unique Values: 4 Type: float64	Missing Values: 50	
GIDP	Grounded into double plays by opposing batter	Unique Values: 7 Type: float64	Missing Values: 50	
Table Name/ Description	*Salaries* Description: Player salary data A data frame with 26428 observations on the following 5 variables			
Variable Name	**Description**	**Statistical Summary/Type**	**Errors**	**Used in final dataset?**
yearID	Year	Min-Max: 1985 – 2016 Unique Values: 32 Type: int64	None	Yes/No
teamID	Team; a factor	Unique Values: 35 Type: Object	None	
lgID	League; a factor	Unique Values: 2 Type: Object	None	
playerID	Player ID code	Unique Values: 5149 Type: Object	None	
salary	Salary	Min-Max: 0 - 33000000 Unique Values: 3393 Type: int64	A few values with 0's; considered as missing values	
Table Name/ Description	*Schools* Description: Information on schools players attended, by school A data frame with 1207 observations on the following 5 variables			
Variable Name	**Description**	**Statistical Summary/Type**	**Errors**	**Used in final dataset?**
schoolID	school ID code	Unique Values: 1207 Type: Object	None	Yes/No
name_full	school name	Unique Values: 1199 Type: Object	None	
city	city where school is located	Unique Values: 856 Type: Object	None	
state	state where school's city is located	Unique Values: 49 Type: Object	None	
country	country where school is located	Unique Values: 1 Type: Object	None	
Table Name/ Description	*SeriesPost* Description: Post season series information A data frame with 343 observations on the following 9 variables			
Variable Name	**Description**	**Statistical Summary/Type**	**Errors**	**Used in final dataset?**
yearID	Year	Unique Values: 123 Type: Object	None	Yes/No
round	Level of playoffs	Unique Values: 14 Type: Object	None	

teamIDwinner	Team ID of the team that won the series; a factor	Unique Values: 42 Type: Object	None	
lgIDwinner	League ID of the team that won the series; a factor with levels AL NL	Unique Values: 3 Type: Object	None	
teamIDloser	Team ID of the team that lost the series; a factor	Unique Values: 45 Type: Object	None	
lgIDloser	League ID of the team that lost the series; a factor with levels AL NL	Unique Values: 3 Type: Object	None	
wins	Wins by team that won the series	Unique Values: 6 Type: int64	None	
losses	Losses by team that won the series	Unique Values: 6 Type: int64	None	
ties	Tie games	Unique Values: 2 Type: int64	None	
Table Name/ Description	*Teams* Description: Yearly statistics and standings for teams A data frame with 2925 observations on the following 48 variables			
Variable Name	**Description**	**Statistical Summary/Type**	**Errors**	**Used in final dataset?**
yearID	Year	Min-Max: 1871 – 2019 Unique Values: 149 Type: int64	None	
lgID	League; a factor with levels AA AL FL NL PL UA	Unique Values: 6 Type: Object	Nulls: 50	
teamID	Team; a factor	Unique Values: 149 Type: Object	None	
franchID	Franchise (links to TeamsFranchises table)	Unique Values: 120 Type: Object	None	
divID	Team's division; a factor with levels C E W	Unique Values: 3 Type: Object	Nulls: 1517	
Rank	Position in final standings	Unique Values: 13 Type: int64	None	
G	Games	Unique Values: 13 Type: int64	None	
Ghome	Games played at home	Unique Values: 38 Type: float64	Nulls: 399	
W	Wins	Unique Values: 112 Type: int64	None	
L	Losses	Unique Values: 114 Type: int64	None	
DivWin	Division Winner (Y or N)	Unique Values: 2 Type: Object	Nulls: 1545	
WCWin	Wild Card Winner (Y or N)	Unique Values: 2 Type: Object	Nulls: 2181	
LgWin	League Champion(Y or N)	Unique Values: 2 Type: Object	Nulls: 28	
WSWin	World Series Winner (Y or N)	Unique Values: 2 Type: Object	Nulls: 357	

R	Runs scored	Unique Values: 618 Type: int64	None	
AB	At bats	Unique Values: 1106 Type: int64	None	
H	Hits by batters	Unique Values: 732 Type: int64	None	
2B	Doubles	Unique Values: 309 Type: int64	None	
3B	Triples	Unique Values: 126 Type: int64	None	
HR	Homeruns by batters	Unique Values: 259 Type: int64	None	
BB	Walks by batters	Unique Values: 572 Type: float64	Nulls: 1	
SO	Strikeouts by batters	Unique Values: 1104 Type: float64	Nulls: 16	
SB	Stolen bases	Unique Values: 322 Type: float64	Nulls: 126	
CS	Caught stealing	Unique Values: 137 Type: float64	Nulls: 832	
HBP	Batters hit by pitch	Unique Values: 97 Type: int64	Nulls: 1158	
SF	Sacrifice flies	Unique Values: 56 Type: float64	Nulls: 1541	
RA	Opponents runs scored	Unique Values: 604 Type: int64	None	
ER	Earned runs allowed	Unique Values: 640 Type: int64	None	
ERA	Earned run average	Unique Values: 392 Type: float64	None	
CG	Complete games	Unique Values: 147 Type: int64	None	
SHO	Shutouts	Unique Values: 32 Type: int64	None	
SV	Saves	Unique Values: 66 Type: int64	None	
IPouts	Outs Pitched (innings pitched x 3)	Unique Values: 830 Type: int64	None	
HA	Hits allowed	Unique Values: 744 Type: int64	None	
HRA	Homeruns allowed	Unique Values: 244 Type: int64	None	
BBA	Walks allowed	Unique Values: 563 Type: int64	None	
SOA	Strikeouts by pitchers	Unique Values: 1130 Type: int64	None	
E	Errors	Unique Values: 455 Type: int64	None	
DP	Double Plays	Unique Values: 199 Type: int64	None	
FP	Fielding percentage	Unique Values: 166	None	

		Type: float64		
name	Team's full name	Unique Values: 139 Type: object	None	
park	Name of team's home ballpark	Unique Values: 217 Type: object	Nulls: 34	
attendance	Home attendance total	Unique Values: 2638 Type: float64	Nulls: 279	
BPF	Three-year park factor for batters	Unique Values: 44 Type: int64	None	
PPF	Three-year park factor for pitchers	Unique Values: 42 Type: int64	None	
teamIDBR	Team ID used by Baseball Reference website	Unique Values: 101 Type: object	None	
teamIDlahman45	Team ID used in Lahman database version 4.5	Unique Values: 148 Type: object	None	
teamIDretro	Team ID used by Retrosheet	Unique Values: 151 Type: object	None	
Table Name/ Description	*TeamFranchises* Description: Information about team franchises A data frame with 120 observations on the following 4 variables			
Variable Name	**Description**	**Statistical Summary/Type**	**Errors**	**Used in final dataset?**
franchID	Franchise ID; a factor	Unique Values: 120 Type: object	None	Yes/No
franchName	Franchise name	Unique Values: 99 Type: object	None	
active	Whether team is currently active (Y or N)	Unique Values: 2 Type: object	Nulls: 25	
NAassoc	ID of National Association team franchise played as	Unique Values: 12 Type: object	Nulls: 108	
Table Name/ Description	*TeamsHalf* Description: Split season data for teams A data frame with 52 observations on the following 10 variables			
Variable Name	**Description**	**Statistical Summary/Type**	**Errors**	**Used in final dataset?**
yearID	Year	Min-Max: 1981 - 1981 Unique Values: 1 Type: int64	None	Yes/No
lgID	League; a factor with levels AL NL	Unique Values: 2 Type: object	None	
teamID	Team; a factor	Unique Values: 26 Type: object	None	
half	First or second half of season	Unique Values: 2 Type: int64	None	
divID	Division	Unique Values: 2 Type: object	None	
DivWin	Won Division (Y or N)	Unique Values: 1 Type: object	None	

Rank	Team’s position in standings for the half	Unique Values: 7 Type: int64	None	
G	Games played	Unique Values: 13 Type: int64	None	
W	Wins	Unique Values: 20 Type: int64	None	
L	Losses	Unique Values: 18 Type: int64	None	