1. atbats
    a. Pk: ab_id
    b. Fk:
        i. g_id -> games.g_id
        ii. batter_id -> players_names.id
        iii. pitcher_id -> players_names.id
    c. comments:
        i. we can combine 2019_atbats and atbats(which has 2015~2018 data) (may cause many issues with empty columns)
        ii. “p_score” means the score for the pitcher’s team
        iii. “top” means whether it is top inning or not
        iv. “o” means outs
        v. “stand” means whether the batter is left or right-handed
        vi. we can probably create two attributes “home_Score” or “away_score” to indicate their current score.
        Or not, we can just use “top” and “p_score” to determine the current score for either team
    d. update March 19
        i. 1NF: passed
        ii. 2NF: passed
        iii. 3NF: passed
        iv. BCNF: passed

2. ejections
    a. Pk: ab_id, player_id but they are foreign keys…
        Or perhaps “des” and “date”?
    b. Fk:
        i. ab_id -> atbats.ab_id
        ii. player_id -> players_names.id
        iii. g_id -> games.g_id
        iv. dates -> games.date
    c. comments:
        i. might need to come up with a proper pk
        ii. not data for 2019
        iii. not sure what “event_num” means. They are linked to event_num in “pitches” table but lots of data are missing so I am not sure if we should still keep this
        iv. BS: ‘Y’ if ejection was for arguing balls and strikes, empty otherwise
        v. CORRECT: if BS ejection is correct. C is correct, I is not
        vi. Note that many of these ejections actually are not related to the players, but rather the coaches and managers. For the sake of our database that is focused on player stats, we will not include these records in our database.
    d. update March 19
        i. 1NF: passed
        ii. 2NF: not passed because pk are ab_id and player_id but g_id depends on ab_id only. This is partial dependency.

1. Remove “g_id”

iii. 3NF: not passed because “date” depends on “ab_id”(ab_id -> g_id -> date) and “is_home_team” depends on “ab_id” and “team” (ab_id -> g_id -> home_team)

1. Remove “date” and “is_home_team”. You can find them in table “games”

iv. BCNF: will pass if 3NF is fixed

v. Remarks: remove “des”?

3. games
   a. Pk: g_id
   b. Fk: none
   c. comments:
      i. we can combine 2019_games and games(which has 2015~2018 data). However, 2019 is missing some data including umpires, wind, elapsed time, attendance, start time, weather, delay
      ii. we may need to discuss how to deal with missing data in 2019. Assigning them to NULL?
      iii. We don’t have a database for teams and what we have now is just a short term of the team like “TOR” which represents the Toronto bluejays. Should we create one just like the players_names?
   d. update March 19
      i. 1NF: not passed. Wind is multivalued
         1. Break it down to wind speed and wind direction
      ii. 2NF: passed
      iii. 3NF: passed
      iv. BCNF: passed

4. pitches
   a. Pk: event_num (they are meant for comparing the data with ejections so I believe they are unique)
      Or we can have ab_id, and pitch_num as pk
   b. Fk:
      i. ab_id -> atbats.ab_id
   c. comments:
      i. we can combine 2019_pitches and pitches(which has 2015~2018 data).
         1. Possible issue as some of these columns have empty data for the 2019_pitches
      ii. Lots of unnecessary data. We can get rid of all the xyz coordinates. We can keep the “start speed” and “end speed” and “spin rate” (but 2019 does not have spin rate).
      iii. “code” and “type” are identical in most cases but we can keep both. “code” can be used if the user is looking for something specific. “type” is just a simplified version of it
      iv. Attributes after “code” are worth keeping
      v. If we don’t want to use event_num then we can get rid of it

    d. update March 19
        i. 1NF: passed
        ii. 2NF: passed
        iii. 3NF: not passed. “type” is a simple version of “code”. “type” depends on “code”
            1. Removed because there are some inconsistencies here between 2015-2018 and 2019
        iv. BCNF: passed

5. players_names
    a. Pk: id
    b. update March 19
        i. 1NF: passed
        ii. 2NF: passed
        iii. 3NF: passed
        iv. BCNF: passed

----Some Improvements we can consider or normalizations (March 8) ----

1. atbats
    a. remove “stand” and “p_throws” and add them to the players_names table instead since they are depending on the players
       Though some players can actually throw/bat left and right but this is very rare
2. games
    a. do we want the umpires? If not, we should remove it. If yes, we should create a table like umpires_names and store all the names of the umpires
       Though I think most of the time we are not interested in the umpires and we are missing this in 2019 anyway
    b. should wind be separated into two columns?
    c. Do we really care about delays?
3. Ejections
    a. “des” column is really bad. Should we remove this?
    b. Remove “event_num”?
    c. Is “is_home_team” necessary?
4. Pitches
    a. Can we make “Code” and “type” better? Should we remove “type”? They are identical most of the time and “type” is just a simplified version of “code”
5. players_names
    a. add players’ bat/throw position L/R
6. teams_names
    a. perhaps we can create a table called teams_names and store all the names of the teams

----Assumption and each tables’ data, changes, keys and constraints (March 19) ----

Assumptions

1. The data is for MLB regular season only. Baseball rules are applied
2. At bat ids and game ids are not related. They are related to the year instead.
3. Players can bat/pitch left or right-handed or both. They can also switch their stand in a game. Therefore, there's no relation between a batter/pitcher and whether they are going to bat/pitch left or right-handed at each at bat.
4. Umpires are not that important for general fans. They are removed from the table because we don't have sufficient information to create another table for umpires. It lacks primary key and we don't have the information for 2019 as well.
5. A team can play in different venues when they are the home team, and we don't have the relations for that. Therefore, there's no specific home field for each team. For instance, TOR's home field is Rogers Centre in Toronto, but they usually play couple games in Montreal each year.
6. Most stats of each pitch are too detailed that most of the general fans may not care or understand. We will focus on the stats that are generally available and visible for the audiences in game or on TV.

BLACK: original data
RED: removed
ORANGE: moved from other tables
PURPLE: new added

1. atbats
    a. inning
        i. 0 < int
    b. top
        i. char(5) in ('TRUE', 'FALSE')
    c. ab_id
        i. pk
        ii. decimal (10)
        iii. <year><serial_num>
        iv. can add some constraints like what we did for A3
    d. g_id
        i. fk from games
        ii. decimal (9)
        iii. <year><serial_num>
        iv. can add some constraints like what we did for A3
    e. p_score
        i. int >= 0
    f. batter_id
        i. fk from players_names
        ii. decimal (6)
    g. pitcher_id
        i. fk from players_names
        ii. decimal (6)
    h. stand
        i. Can rename to "bat_dir"
    i. p_throws

        i. Can rename to “pitch_dir”
    j. event
        i. char(25)
    k. outs_after
        i. 0 <= int <= 3
        ii. Can rename to “outs after” as it indicates the number of outs after this at bat. It’s more clear

2. games
    a. attendance
        i. int
    b. away_final
        i. int >= 0
    c. away_team
        i. fk from teams_names
        ii. char(3)
    d. date
        i. datetime
    e. elapsed_time
        i. int
    f. g_id
        i. pk
        ii. decimal (9)
        iii. <year><serial_num>
        iv. can add some constraints like what we did for A3
    g. home_final
        i. int >= 0
    h. home_team
        i. fk from teams_names
        ii. char(3)
    i. start_time
        i. time format
    j. umpire_1B
    k. umpire_2B
    l. umpire_3B
    m. umpire_HP
        i. I don’t really like them because officials should not be the reason who impacts the game. Therefore, they are not usually the ones we care. However, if we really need this, then we probably need to have a new table umpires_names and create an id for all of them…
    n. venue_name
        i. char(64)
    o. weather_degrees
        i. int
        ii. get rid of degrees
    p. wind

        i. multivalued. We separate this into speed and condition
    q. wind_speed_mph
        i. int
    r. wind direction
        i. char(15)
    s. delay
        i. int

3. Ejections
    a. ab_id
        i. pk with player_id
        ii. fk from atbats
        iii. decimal (10)
        iv. <year><serial_num>
        v. can add some constraints like what we did for A3
    b. des
        i. char(100)
        ii. renamed to description
    c. player_id
        i. pk with ab_id
        ii. fk from player_names
        iii. decimal (6)
    d. BS
        i. char(5) in ('TRUE', 'FALSE')
        ii. Originally it's just Y and NULL but we can change it to TRUE and FALSE
        iii. Can rename to argue_ball_strikes
    e. CORRECT
        i. char(5) in ('TRUE', 'FALSE')
        ii. Originally it's just C and I but we can change it to TRUE and FALSE
        iii. Can rename to correct_ejection
    f. team
        i. fk from teams_names
        ii. char(3)
        iii. make it upper letter
    g. date
        i. violate 3NF
    h. event_num
        i. useless
    i. is_home_team
        i. violate 3NF
    j. g_id
        i. violate 2NF

4. Pitches
    a. px

b. pz
c. start_speed
    i. float
d. end_speed
    i. float
e. spin_rate
    i. float
f. spin_dir, break_ang, break_leng…
    i. they are too detailed that most of the general fans don't usually care
g. code
    i. char(2) in (B, *B, S, C, F, T, L, I, W, M, P, Q, R, X, D, E, H, V, Z)
h. type
    i. Removed because they are not that useful
i. pitch_type
    i. char(2) in (CH, CU, EP, FC, FF, FA, AB, FO, FS, FT, IN, KC, KN, PO, SC, SI, SL, UN)
j. event_num
    i. useless
k. b_score
    i. int >= 0
l. ab_id
    i. pk with pitch_num
    ii. fk from atbats
    iii. decimal (10)
    iv. <year><serial_num>
    v. can add some constraints like what we did for A3
m. b_count
    i. 0 <= int <= 4
    ii. can rename to ball_count
n. s_count
    i. 0<= int <= 2
    ii. can rename to strike_count
o. outs
    i. 0 <= int < 3
p. pitch_num
    i. pk with ab_id
    ii. int > 0
q. on_1b
    i. char(5) in ('TRUE', 'FALSE')
r. on_2b
    i. char(5) in ('TRUE', 'FALSE')
s. on_3b
    i. char(5) in ('TRUE', 'FALSE')

5. players_names
    a. id

        i. pk
        ii. decimal (6)
    b. first_name
        i. char(15)
    c. last_name
        i. char(20)

6. teams_names
    a. abbreviation
        i. pk
        ii. char(3)
        iii. make the name upper letter
    b. city
        i. char(15)
    c. short_name
        i. char(15)