#### 1. Discussion of how you converted the dataset into tables.

From the csv file I downloaded from baseball reference, I manually filtered out the overlapping columns and group them into tables to minimize data redundancies. For instance, instead of keeping the player names in each category (batting, pitching, fielding), I have created 3 sperate tables for players. Same happens for teams and managers.

In MySQL, I wrote the create statements to create tables under a new schema, and I manually set the primary and foreign keys within the queries. Due to the large amount of data, I used Mr. Data Converter (thdoan.github.io) to assist me while inserting the data. It generates the insert statement, and all I have to do is to make sure they are inserting to the correct table.

# 2. Challenges faced during importing of your data and how did you overcome these data importation challenges.

I ran into a couple challenges during the process. First being renaming the columns, some column names contain special characters such as "+", and some start with numbers. To create tables in MySQL, these kinds of column names need to be changed, or else it is going to return errors. Therefore, I must change "+" into plus and put an underscore before columns whose name starts with a number.

I also ran into an error while inserting data for the table. It returns the following: Error Code: 1452. Cannot add or update a child row.

Later I am able to notice that in the player\_batting, player\_pitching, and player\_fielding tables, the original csv contains an overall row for players who changed team during the season. So, when a player played for team A and B in the same season, there are 3 rows of records, being stats from Team A, stats from Team B, and Total stats that combine the two rows. The problem is the data has this kind of rows labeled as "TOT" in the TeamID column, which cannot be connected to a TeamID in the Teams table. Therefore, I manually filtered out all the TOT rows and deleted them.

## 3. A complete data dictionary for every table in your database.



#### 4. Business Questions

- 1. Which player hit the most HR for each team?
- 2. Who is the most frequently traded player in the 2023 season?
- 3. Who has the highest hitting percentage in each age level? List out the leaders and the average of each age level.
- 4. Does successful rate of challenges by the managers translate to wins? List out number of wins and calculate the rate.
- 5. Does lower Batting Park Factor really tie with lower batting averages?
- 6. Do more fielding errors lead to more losses?
- 7. List out each team's Runs per game and Runs allowed per game. How does this margin impact the result of games?
- 8. Does Runners Left On Base impact Runs scored per game? List these two columns out for each team.
- 9. Which player(non-pitcher) played the most game not starting?
- 10. List out any player who has appeared as both batter and pitcher and their stats.
- 11. Ranking pitchers who created the most double plays.

## **Proof of Table Creation**



















