ETL Project Analysis Report

Lisa Weinstein, Connor Berek, Ryann Green

- **The sources of data that you will extract from.**

  The Kaggle database was used to obtain csvs about baseball teams over the years. Some other data we collected was on players who made it to the all-star game, as well as batting and pitching statistics on players. The goal of this project was to determine the best team in the history of baseball.

- **The type of transformation needed for this data (cleaning, joining, filtering, aggregating, etc).**

  For all of the datasets, there were columns included that were not relevant to the analysis for the best player/team of all time so these columns were first removed from the datasets within Jupyter Lab. Afterwards, we noticed that there were null or "na" values in some of the cells so we replaced those values with zeros for int data types and "None" for string data types to avoid errors when it was time to load the data into a SQL database. Next we renamed some of the column headers out of caution just in case some of the column names would conflict with reserved words when needing to create the tables in our database (ex. "year" to "year_" and "double" to "double_"). After we had finished cleaning the csv files, we established a connection to the local postgres server, created an engine to funnel the cleaned data into its corresponding table in the database that was created.

- **The type of final production database to load the data into (relational or non-relational).**

  We utilized postgresql, specifically PGAdmin which is a relational database, for the purposes of this assignment because we are utilizing csv files that necessitate organization due to the csv data and its dependencies on one another. Furthermore, we utilized the Quick Database Diagram app that was shown to us in order to better see the relationships between these datasets that were used. (Ref: https://app.quickdatabasediagrams.com/)

- **The final tables or collections that will be used in the production database.**

  After we finished cleaning the csv's and making them ready to be transferred into the database we created, baseball. In this database, there are 5 tables: Team, Allstar, Batting, Pitching and Player. Each one of these tables has a connection to its corresponding csv file that was cleaned utilizing JupyterLab. When creating the tables, I made sure to not allow NULL values when loading the cleaned csv files into the table in order to catch any missed null values during the cleaning process. We did, however, allow null values for dates due to the inconsistency across the formatting of both the "debut" and "last game" columns of the players csv. If given more time, we would have been able to better hash out this issue. This data, once it was loaded into the database allows any baseball fanatic to really get down to the roots of the best franchise, the best players, the best teams, year to year as I demonstrate in some of the joins performed on the query.sql file of our Repository.