Team Beta: Steve Kennedy, Camilla Inhapim, Michael Goff, Jason Hanlin, Marlon Ruiz

Topic: Major League Baseball

Task: ETL

Team Beta extracted data from sources with a focus on Major League Baseball statistics. The repository and source links for the project are as follows:

* <https://github.com/tenntully/Project2-TeamBeta.git>
* <https://www.seanlahman.com/baseball-archive/statistics/>

Please see code for extraction operation as follows:

Graphical user interface, text, application, email

Description automatically generated

**Figure 1.** VS code representation of extraction process code via pandas in jupyter notebook.

After extraction, two data sets, one for teams and one for franchises, were read into pandas data frames. Then, the transform process began with a left join on franchise ID to create a single database. Numerical values were formatted as floats and redundant/duplicate columns were dropped. Columns were renamed to provide clarity on the meaning of descriptive labels as follows:

A picture containing table

Description automatically generated

**Figure 2.** VS code representation of transform process code via pandas in jupyter notebook.

Next, the load operation was performed with sqlalchemy library. The database was loaded into PostgreSQL. This process began with a session query to connect jupyter notebook to the data frame. Then, the ORM session was established to provide an interface for forthcoming queries in PostgreSQL. The team created the desired table with the 9 renamed columns. Please see loading operation code and resulting table in figures 3 and 4 respectively as follows:

Graphical user interface, text, application

Description automatically generated

**Figure 3.** VS code representation of load process code via pandas in jupyter notebook.

Graphical user interface, application, table

Description automatically generated

**Figure 4.** PostgreSQL table formed from load process.