ETL Project

* Due: 5/14/19
* Members: Kent Hazen, Amer Lukaesko, Ryan Tamashiro, Julie Kiernan
* Data Sources: Kaggle:

Project Proposal:

* Create database on MLB baseball player information from kaggle
* CSV files:
  + https://www.kaggle.com/wduckett/moneyball-mlb-stats-19622012/downloads/moneyball.zip/1
  + https://www.kaggle.com/freshrenzo/lahmanbaseballdatabase/downloads/lahmanbaseballdatabase.zip/1

Project Report:

* + **E**xtract: Kaggle databases pulled (manually pulled salary, player, player stats, and batting statistics data; formatted as .csv)
  + **T**ransform:
  + Imported data using pandas in jupyter notebook
  + Data sets relatively clean; eliminated columns deemed unnecessary for this project and rows where playerID was missing
  + cleaned data converted to dataframes; then into lists of dictionaries to utilize in mongo db
  + **L**oad: final database; MongoDB (non-relational database)
  + Database = baseball
  + Collections= player\_data; batting\_data
  + Mongo chosen as final data schema because of its flexibility and the multiple data values available to query