Project 2: Extract, Transform, Load

Jasmine Jones, Haydn Whitmyer, Nicholas Noark

Basketball Data

Extraction:

- Data taken from a SQLite database found at https://www.kaggle.com/wyattowalsh/basketball
- Statistics for the 2020-21 NBA season scraped using Pandas from basketball-reference.com
- Data pulled from https://rapidapi.com/api-sports/api/api-nba API

Transformation:

- From the SQLite database, active players pulled from the Players table and joined with the Draft table to create our active_info table.
- Player salaries for the 2020-21 NBA season pulled from the Player_Salary table from the SQLite database.
- Per 100 possessions player statistics for the 2020-21 NBA season scraped from basketball-reference and put into a Pandas dataframe and exported to CSV
- Per 36 minutes player statistics for the 2020-21 NBA season scraped from basketball-reference, rate stats that overlap with per 100 possessions data removed, placed into a Pandas dataframe, an exported to CSV
- From API-NBA, pulled gameDetails, players/playerId, and teams/teamId to create tables containing select info about each game, player, and team.
    - Converted data from JSON format to a Pandas dataframe, then exported each to CSV

Load:

- PostgreSQL database, “mybasketball”, created using the psycopg2 package in Python
- Initial tables created using pgAdmin4 query tool
- Data loaded into corresponding tables using pandas and SQLAlchemy packages