Diana Andrade

Samuel Slomowitz

Yarely Vargas

Matthew Witschorke

**NCAA March Madness ETL Using Beautiful Soup and PostgreSQL**

NCAA March Madness is a data-driven process of collecting and interpreting information to optimize better outcomes for the tournament. Ken Pomeroy is a venerable analyst and data collector for the annual NCAA bracket. Warren Buffet offers a multi-million-dollar prize to anyone who submits a perfect bracket. To facilitate better predictions and build towards a larger data science project, we begin with an ETL, or Extract, Transform, and Load, database repository. Thus, a thorough data collection method will enable future machine learning and web visualization endeavors.

Our database design contains a relational database that hinge on primary and foreign keys, especially a “team id,” which was simply the rank of the team in the NCAA tournament. Given that 68 team now play in the annual NCAA bracket, we assign 68 id to each team, respectively. Coaches we rank not on their coaching expertise and perform but rather on their basketball career in high school and/or college. Player statistics included Points per Game, Assist, Rebounds, Steals, and 3-pointers. The TV Schedule included Round of the tournament, Day of the Game, time, TV network, city, and venue.

Coaches Rank was web scraped from ESPN using Beautiful Soup. Data cleaning included removing non-text items, unneeded punctuation, and splitting apart rank, first name, and last name. The TV Schedule including the outcome of the first 44 games of the tournament was web scraped from NCAA.com using Splinter. We set up a complicated yet effective for-loop to loop through the rows of a table to pull game team information, time, and TV network data. Player statistics were web scraping from Real-GM using Splinter, pandas, and Beautiful Soup. Team statistics were web scraped from Real-GM using pandas.