Diana Andrade
Samuel Slomowitz
Yarely Vargas
Matthew Witschorke

NCAA March Madness ETL Using Beautiful Soup and PostgreSQL

Introduction

NCAA March Madness is a data-driven event that entails 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 (Extract, Transform, and Load) database repository. Thus, a thorough data collection method will enable future machine learning and web visualization endeavors.

Database Design

Our database design contains a relational database that hinges on primary and foreign keys, especially a “team id,” which was simply the rank of the team in the NCAA tournament. Given that 68 teams now play in the annual NCAA bracket, we assign 68 ids to each team, respectively [this is found in the teams table]. Coaches in the coaches table are ranked not on their coaching expertise and performance but rather on their basketball career in high school and/or college. Player statistics table includes points per game (PPG), assists (APG), rebounds (RPG), steals (SPG), and 3-pointers (3P%). The TV schedule table includes the round of the tournament, day of the game, time of the game, TV network, city, and venue.

Extracting Data via Web Scraping

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 iterate through the rows of a table to pull game team information, time, and TV network data. Player statistics were web scraped from Real-GM using Splinter, pandas, and Beautiful Soup. Team statistics were web scraped from Real-GM using pandas.

Analysis

Utilizing PGAdmin, we ran five SQL queries to analyze the data. The queries revealed that the Second Round of the tournament had a top 10 seed player versus a seed greater than 13 in rank. Coaches Mark Few from Gonzaga and Tommy Lloyd from Arizona coached high performing players in the points per game category. Arizona, Duke, Illinois, and Iowa State, all respectable teams in NCAA tournament history, had primetime showing on TV Networks in the First Round. Tyrese Hunter from Iowa State, J.D. Notae from Arkansas, and Charlie Moore from Miami were solid all-around players that top the charts in terms of steals per game, rebounds per game, and points per game while being featured in high seeded teams in the tournament.

Using SQLAlchemy we performed a more thorough analysis of the data sources we collected. The first data visualization featured the top ten teams based on the players' points per game. This seaborn plot revealed that Tennessee had the top player points per game while Purdue had the lower player points per game. We constructed a donut chart that represented the top performing players in the category of assists. Tyler Kolek from Marquette escalated to the top of assist category in the first round and part of the second round of March Madness. Finally, we

looked at average total game score by Time and TV Network. This revealed that it's better to watch CBS for higher scoring games in the tournament. Night time games on Saturday are typically higher scoring as well.

Conclusion

Some future work includes automating the transformation and loading of data from Python to a PostgreSQL database. One difficulty we encountered in web scraping included that the data sources changed constantly, and the website changed from day to day. However, our bottom line is that since Gonzaga and Arizona have the high metric of points per game from the season data, both should be favored to win the tournament.