Ethan Pruzhansky

Reflection

As a long-time hockey enthusiast, I decided to analyze statistics for the top 250 NHL players to determine their contribution levels during games. The web crawling process was a fascinating challenge. Initially, I struggled to write the correct code for extracting the required data pattern. The complication arose from the website featuring two tables with identical patterns. My concern was about extracting only the necessary table. Fortunately, the links in both tables were listed in descending order, and the table I needed contained 251 entries. Thus, I used .head(251) to collect the links I required.

The web scraping phase was relatively straightforward. I utilized the pointer from the inspect element feature to locate the table, then applied soup.find to record the specific classes needed to extract the HTML element. Subsequent data cleaning involved using regular expressions to refine the dataset. A notable challenge was combining two tables with identical column names and statistics. To resolve this, I renamed the columns and discarded the redundant one.

The analysis aimed to identify which hockey players contributed most significantly in each game. This required calculating the number of games each player participated in per year and their points scored per game. I then established a correlation between points and games, using a line graph to display the performance of the top 100 and bottom 100 players from my dataset.