Ryan Gilbert
July 28th, 2025
CMP 262- Project 2 Part 1

Project 2

Objective: In this project, you will be scraping data from a website and then performing an analysis. You must scrape data from 5 different web pages.

Review the list of examples of scrapeable websites in this module. Explore these sites or other known scrapeable sites and finalize the questions you may have about one of these datasets.

Requirements:

1. Name and the URLs (at least 5) of the site you will be scraping data from.
2. At least five questions that you'd like to explore from your selected dataset. You will ultimately pick 4 questions and at least two should use a plot to answer the question. Your questions should not be simple answers that are obvious by just looking at the web page. A good data science question is one that performs some type of summary calculation on the data from the website.
3. Plan for data scraping (e.g., pull data from the table with id=xxx from five different pages that are named by year/team-name)
4. Restrictions noted in the Robots.txt file (state if there are no restrictions in the robots.txt file)

Summary: For this project, the website Hockey Reference will be used for data scraping and analysis. After exploring other hockey sites (NHL.COM, MoneyPuck), those sites either had more restrictive access for webscraping and/or would not be easily scraped using BeautifulSoup/Python because their stat pages are JavaScript rendered.

Ryan Gilbert
July 28th, 2025
CMP 262- Project 2 Part 1

Deliverables

1) **Names & URLS**
    a) **2024-2025 NHL Skater Statistics:** https://www.hockey-reference.com/leagues/NHL_2025_skaters.html
    b) **2024-2025 NHL Goalie Statistics:** https://www.hockey-reference.com/leagues/NHL_2025_goalies.html
    c) **New Jersey Devils Skaters:** https://www.hockey-reference.com/teams/NJD/skaters.html
    d) **New Jersey Devils Won-Loss Records:** https://www.hockey-reference.com/teams/NJD/head2head.html

2) **Data Questions**

    a) Of the Top 50 Goal Scorers from the 2024-2025 NHL Season, what Percentage of their Goals Scored come from Even Strength Goals vs from the PowerPlay?
    b) For those goalies who started the majority of games for their team (41+) in the 2024-2025 season, who were the top five goalies in Goals Saved Above Expected? And did this translate into those same goalies receiving the most Vezina Trophy votes?
    c) Who were the Top 20 Most Productive (Points Per Game) players in NJ Devils franchise history (who played at least 200 games).
    d) Which 5 teams do the Devils have the greatest average goal differential (Average Goals/Game - Goals Against/Game) of all time and are those the same 5 teams with the highest Points Percentage against?

3) Data Scraping

For Question A, the data would be scraped from the first link “2024-2025 NHL Skater Statistics”. Once the data is pulled from the site, the columns “G”, “EVG” and “PPG” would be used to answer the data question. For Question B, the data would be scraped from the second link, “2024-2025 NHL Goalie Statistics”, where those rows with “GP” of over 41 would be used along with the “GSAA” and “Awards” columns to answer the data question. Question C would be answered using both the data from the “2024-2025 NHL Skater Statistics” webpage and the “2025 Stanley Cup Playoff Skater Statistics” webpage. For both, PPG would be calculated by dividing “P” by “GP”.The fourth data question would be answered by the “New Jersey Devils” skaters webpage, and the fifth would be answered using the data included in the “New Jersey Devils Won-Loss Records” page.

4) **Restrictions from Robots.txt File**

- **AhrefsBot** and **GPTBot** are completely blocked from accessing the site.
- **TwitterBot** is allowed full access (no restrictions).
- **All other bots** are restricted from accessing specific parts of the site, including:
    - Hockey-related content (`/hockey/`)
    - Game logs, splits, scoring pages
    - Player search and certain boxscore pages
    - User-related or private areas (`/my/`, `/req/`, `/short/`, etc.)
- A **3-second crawl delay** is required for all general bots.

Works Cited

OpenAI. (2025, July 29). *ChatGPT (July 29 version)* [Large language model]. https://chat.openai.com/chat