Top Spin

Project Proposal
CIS 5500: Database
University of Pennsylvania

Members (email; github):

- Akanksha Ashok (aashok@seas.upenn.edu; @aashok12)
- Noah Capp (@noahcapp@seas.upenn.edu; @noahmcapp)
- Bilal Ali (bilalali@seas.upenn.edu; @bilala45)
- Peter Akioyamen (peterai@seas.upenn.edu; @peter-ai)

Project Description

Top spin is a web application that allows both tennis fanatics and casual fans to gain insight into the most infamous question in any given sport - who is the greatest of all time (GOAT). It aggregates data about athletes, rankings, and matches dating back to 1968, and match odds from 2000 to 2019. The web app will display various dynamic summary views and more in-depth pages about historical tennis players and matches. Aligned with this, it will provide users the functionality to create matches with players from different eras of tennis, and show the user who was more likely to win given the user defined parameters about the match.

Datasets

The data for the project will be gathered from three github repos. Namely one for the ATP tennis circuit (men's), another for the WTP (women's circuit), and a third which has betting odds of matches. Data will be extracted from several cleaned CSV files, pre-processed and stored in a database. Data can be found in the following repos:

Relevant Links

- ATP tennis (https://github.com/JeffSackmann/tennis_atp)
- WTP tennis (https://github.com/JeffSackmann/tennis_wta)
- ATP & WTP betting odds (https://github.com/chief-r0cka/MLT)

Queries

Query for player performance:

Aggregate number of matches played, wins and losses in both single and doubles matches, along with key player performance metrics (aces, faults, etc.) for each player over time.

Query for tournament match results:
Display match data (players, player seeds, match duration, sets won by each player, etc.) for each match in a tournament over time.
Query for head-to-head matchups with previous matches:
Retrieve matches where players have previously played against each other (at least 5 times) over time and across different tournaments to determine which player will have a historical edge in future matches.

Query for head-to-head matchups (predictive)::
Retrieve player performances over time and across different tournaments (must have played at least 5 matches) to determine which player would win in a hypothetical matchup.

Query for matches with high odds discrepancies (David vs Goliath):
Join betting odds data with match data for matches with high odds discrepancies (one player is highly favored) and calculate win percentage for match favorite and underdog to determine players with consistent upsets.

Query for player winning streaks:
Calculate longest consecutive number of wins (winning streaks) of players over time across both male and female competitions.

Query for betting favorites:
Join betting odds data with player performance data and calculate a profitability assessment for players to inform future betting strategies (players that are consistently safe bets, players that perform well as underdogs, etc.)

Summary Statistics (see below table)

- ATP match data has 191303 rows and 49 columns (42.2MB)
  - 61639 distinct athletes
  - Matches from December 1967 to August 2023
- WTA match data has 154870 rows and 49 columns (86.4MB)
  - 65607 distinct athletes
  - Matches from December 1967 to August 2023
- ATP odds data has 54908 rows and 55 columns (12.6MB)
- WTA odds data has 32053 rows and 55 columns (7.4MB)

Summary data description:

- **w_ace**: winner's number of aces
- **w_df**: winner's number of doubles faults
- **w_svpt**: winner's number of serve points
- **w_1stIn**: winner's number of first serves made
- **w_1stWon**: winner's number of first-serve points won
- **w_2ndWon**: winner's number of second-serve points won
- **w_SvGms**: winner's number of serve games
- **w_bpSaved**: winner's number of breakpoints saved
- **w_bpFaced**: winner's number of breakpoints faced
- **l_ace**: loser’s number of aces
- **l_df**: loser’s number of doubles faults
- **l_svpt**: loser’s number of serve points
- **l_1stIn**: loser’s number of first serves made
- **l_1stWon**: loser’s number of first-serve points won
- **l_2ndWon**: loser’s number of second-serve points won
- **l_SvGms**: loser’s number of serve games
- **l_bpSaved**: loser’s number of breakpoints saved
- **l_bpFaced**: loser’s number of breakpoints faced

Statistic	w_ace	w_df	w_svpt	w_1stIn	w_1stWon	w_2ndWon	w_SvGms	w_bpSaved	w_bpFaced
Winner									
mean	6.526239	2.733728	78.128976	47.657888	35.933390	16.729124	12.409395	3.526488	5.164499
std	5.337322	2.364958	29.539587	19.242839	13.852961	6.983693	4.128099	3.086376	4.062636
min	0	0	0	0	0	0	0	0	0
25%	3	1	56	34	26	12	9	1	2
50%	5	2	73	44	33	16	11	3	4
75%	9	4	94	58	43	21	15	5	7
max	113	26	491	361	292	82	90	24	34
Loser									
mean	4.840857	3.484673	80.967071	48.089749	31.955067	14.983604	12.209736	4.813242	8.739995
std	4.680657	2.620183	29.471466	19.402443	14.462356	7.215126	4.136927	3.276748	4.134577
min	0	0	0	0	0	0	0	-6	0
25%	2	2	59	34	22	10	9	2	6
50%	4	3	76	45	30	14	11	4	8
75%	7	5	97	58	40	19	15	7	11
max	103	26	409	328	284	101	91	28	38