Data Construction Project

Find the data at the following link: https://github.com/JaydenCruz2004/Sports-Analytics-Project

The level of badminton analytics data is, in fact, quite limited compared to more popular sports. The BWF World Tour and sites such as badmintonstatistics.net provide career and seasonal statistics, but available data is limited to aggregate match and point numbers—there are no publicly available datasets containing biometric data for players, rally-level data for events, or match-level data that could easily be merged with player information. The SCORE Network dataset (Smith, 2024), originally obtained from badmintonstatistics.net, tracks cumulative World Tour performance from 2018 to 2023 for 185 players across nine variables: player name, discipline (Singles/Doubles), matches played, wins, losses, points scored, points conceded, win percentage, and shot percentage. While this is a good start, it did not include any measure of efficiency scaled for play volume or a means to distinguish players by level of play experience, which would be helpful in comparing players who have competed vastly different numbers of matches over the course of the five years.

To enrich the dataset, three derived variables were computed entirely from the existing columns. First, pt_differential (pts_for − pts_agst) was created to capture each player's overall net point margin. Second, pt_differential_per_match (pt_differential / matches_played) standardizes that margin on a per-match basis, making meaningful comparisons possible between a player with 3 matches and one with 128. Third, experience_tier classifies players as Low (1–5 matches), Moderate (6–20), High (21–50), or Elite (51+) to allow group-level comparisons across competitive exposure. Columns were renamed to consistent snake_case and floating-point variables were rounded. No missing or duplicate values were present. One potential analysis using only the original CSV would be a simple linear regression of pts_for predicting wins the SCORE Network page itself poses this question which would let you quantify how much each additional point scored translates into expected wins, and whether that relationship differs between Singles and Doubles players (testable via an interaction term or separate models).

BWF World Tour Dataset – Data Dictionary

Dataset: badminton_wrld_tour.csv

Observations: 185

Variables: 12

Season Coverage 2018 – 2023

Source: SCORE Sports Data Repository – Smith, A. (2024) 2018 – 2023 Badminton World Tour Points Head to Head

Variable Definitions

The table below defines all 12 variables. The first 9 are drawn directly from the original dataset, the final 3 were derived from those variables.

Variable Name	Type / Units	Source	Definition
player_name	Text	SCORE Network (original)	Full name of the player competing in the BWF World Tour.
discipline	Categorical (Singles / Doubles)	SCORE Network (original)	The event category the player competed in: Singles (individual) or Doubles (pairs).
matches_played	Integer (count)	SCORE Network (original)	Total number of matches the player played in the World Tour between 2018–2023.
wins	Integer (count)	SCORE Network (original)	Total number of matches won by the player between 2018–2023 in the World Tour.
losses	Integer (count)	SCORE Network (original)	Total number of matches lost by the player between 2018–2023 in the World Tour.
pts_for	Integer (points)	SCORE Network (original)	Total rally points scored by the player across all World Tour matches (2018–2023).
pts_agst	Integer (points)	SCORE Network (original)	Total rally points allowed (conceded) by the player across all World Tour matches (2018–2023).

win_pct	Numeric (proportion, 0–1)	SCORE Network (original)	Cumulative win percentage: wins / matches_played. Rounded to 4 decimal places.
shot_pct	Numeric (proportion, 0–1)	SCORE Network (original)	Cumulative shot percentage: pts_for / (pts_for + pts_agst). Represents the share of total rally points won by the player. Rounded to 4 decimal places.
pt_differential	Integer (points)	Derived from pts_for & pts_agst	Net point margin: pts_for minus pts_agst. Positive values indicate the player scored more points than they conceded overall.
pt_differential_per_match	Numeric (points / match)	Derived from pt_differential & matches_played	Average point margin per match: pt_differential / matches_played. Rounded to 2 decimal places. Standardizes point margin for players with different match totals.
experience_tier	Categorical (ordinal)	Derived from matches_played	Ordinal classification of competitive volume based on matches_played: Low (1–5), Moderate (6–20), High (21–50), Elite (51+).

Data Sources

Smith, A. (2024). 2018–2023 Badminton World Tour points head-to-head. SCORE Sports Data Repository. https://data.scorenetwork.org/badminton/badminton_worldtour_2018-23.html

Badminton Statistics. (2023). Player head-to-head statistics. https://www.badmintonstatistics.net

Badminton World Federation. (2023). BWF World Tour. https://bwfworldtour.bwfbadminton.com/

Cleaning & Transformation Notes

- Renamed all columns to consistent snake_case: Player converted into player_name, Matches into matches_played
- win_pct and shot_pct rounded to 4 decimal places for consistency.
- Computed pt_differential = pts_for – pt_against (net point margin across all matches)
- Computed pt_differential_per_match = pt_differential / matches_played, rounded to 2 decimal places. This standardizes the margin metric across players with very different match totals
- Computed experience_tier as an ordinal categorical variable from matches_played: Low (1-5 matches), Moderate (6-20), High (21-50), Elite(51+). Thresholds chosen to reflect meaningful difference in World tour exposure.
- No missing values were present in the original or final dataset. No duplicate player rows were found
- Row index from original CSV was dropped; rows are indexed 0 – 184 in the output file