Badminton is a racket sport played by pairs (*Doubles*) or individuals (*Singles*). Badminton takes place on a court with a net in the middle separating the opponents. Players use rackets to hit a feathered ball known as the birdie across the net. Matches are won through the best of three games, each match is won by the first team to get 21 points. Points are won when the birdie hits the ground in the opposing side's court.

While often considered a recreational sport, it is also played at the professional and Olympic level. The BWF Badminton World Tour is a tournament series with opens hosted in different countries around the world. The tour has been running since 2018 and even took place in 2020 despite the covid-19 pandemic. Statistics have been compiled from each

World Tour, a notable statistic being the *Points H2H* which looks at the points for (*pts\_for*) a player and the points against (*pts\_agst*) them.

Through statistical analysis, patterns can be identified in this [badminton dataset](badminton.csv) (badminton.csv). Investigating the distribution of the *Wins* variable will highlight any patterns and what factors might impact winning in badminton at the World Tour Level.

|  |  |
| --- | --- |
| **Variable** | **Description** |
| *Player* | The name of the player competing. |
| *Category* | The category of competition the player competed in, either *Singles* or *Doubles*. |
| *Matches* | The total number of matches the player played in between 2018-2023 in the World Tour. |
| *Wins* | The total number of wins the player had between 2018-2023 in the World Tour. |
| *Losses* | The total number of losses the player had between 2018-2023 in the World Tour. |
| *pts\_for* | The total number of points scored by the player between 2018-2023 in the World Tour. |
| *pts\_agst* | The total number of points allowed by the player between 2018-2023 in the World Tour. |
| *win\_pct* |  |
| *shot\_pct* |  |

1. Create a histogram of *Wins* and describe the distribution.
2. Obtain the summary statistics of *Wins* and fill them in below.

**Mean:**   **Standard Deviation: Minimum:**

**Maximum:**  **Median:**

1. Calculate the range for *Wins* and for *pts\_for*.
2. The ranges for *Wins* and *pts\_for* are both quite large. What do you think is a possible reason for this flaw in the data?
3. Create a side-by-side box plot of *Wins* for each *Category*. Describe the distribution.
4. Perform a t.test to determine if there is a significant difference in the mean number of *Wins* for *Doubles* vs. *Singles*.

1. Fit a simple linear model of *pts\_for* predicting *Wins*.
   1. Interpret the slope of the model you just fit.
   2. If a player has 216 points calculate their predicted *Wins*.
   3. Assess the overall fit of the model.
   4. Find the 90% confidence interval of the slope of the model.
2. Do you think that *pts\_for* is really the most effective predictor of *Wins*? Why or why not? What could be a better predictor?