Model takes in two team stats, and then returns the predicted winning team.

Accuracy is based on whether the model correctly picked the winning team given the two team stats.

* EDA should explore what correlates to a team winning (win %). (maybe explore certain features with the total amount of wins a team has in a season).
* Data input should be team 1 stats, team 2 stats, and the winning team as the output (and maybe score differential).
* Find which features are correlated to each other, get rid of them

**Part 1 Kruskal-Wallis/Pearson’s on Features:**

* Graph every stat (numerical) vs round reached in tournament (categorical)
  + This would be by year, one year at a time
  + Violin plots (or box plot 😊)
  + Kruskal-Wallis test for significance
* Graph every stat (numerical) vs historical tournament win % (numerical)
  + This would be across years
  + Scatter plot
  + Pearson’s correlation coefficient
* This is to determine which features might predict labels
* USE BONFERRONI
* Explain that we’re trying to understand the relationship/significance with features, so we use all the data. This is in comparison to if we were doing regression for prediction, where we would do a train-test split
  + We did not standardize the data, so the coefficients represent the actual units of the features, not relative importance.
* TODO: If we want to drop features that are similar, do PCA on them and take the first principal component.
  + For example, we don’t need every committee’s ranking, we could combine them using PCA and say that is the “ranking vector”
* TODO: Handle outliers and null values and give a reason/explanation for why we handled them like this

**Part 2: Use Logistic Regression to Determine Important Features**

* Create a logistic regression model with all the features
* Remove one feature, check results, determine if that feature mattered or not
* Make sure the data meets the assumptions of logistic regression
  + There is a linear relationship between the predictor variables and the logit of the response variable. This assumption can be tested using a Box-Tidwell test.
  + **Sample size is large enough**
  + There is no severe multicollinearity among predictor variables. It is assumed that none of the predictor variables are [highly correlated](https://www.statology.org/multicollinearity-regression/) with each other.
    - Do PCA if they are correlated
  + No extreme outliers
  + The observations are independent. It is assumed that the observations in the dataset are independent of each other. That is, the observations should not come from repeated measurements of the same individual or be related to each other in any way.
  + **The response variable is binary**

**Part 3: DNN Predicts Matchups (kinda optional, but not really)**

* Give a DNN two vectors of all the stats, one for each team
* Have it output the winner (binary classification) and the score differential (continuous value)

Feature Ideas:

* Compare win rate to home/away/neutral
* Compare win rate to distance from home campus
  + Or time zones crossed?