ATP Tennis 2000-2019

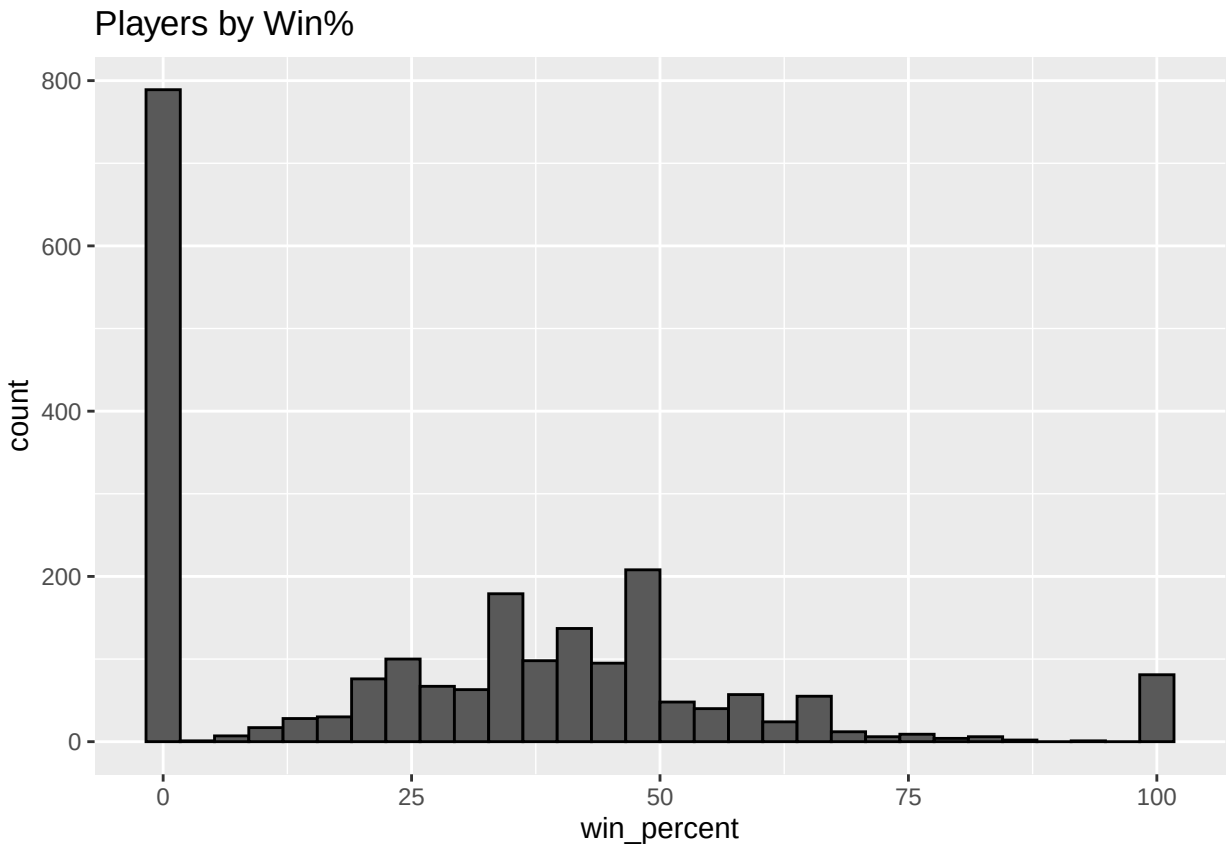
Maxwell Vestrand

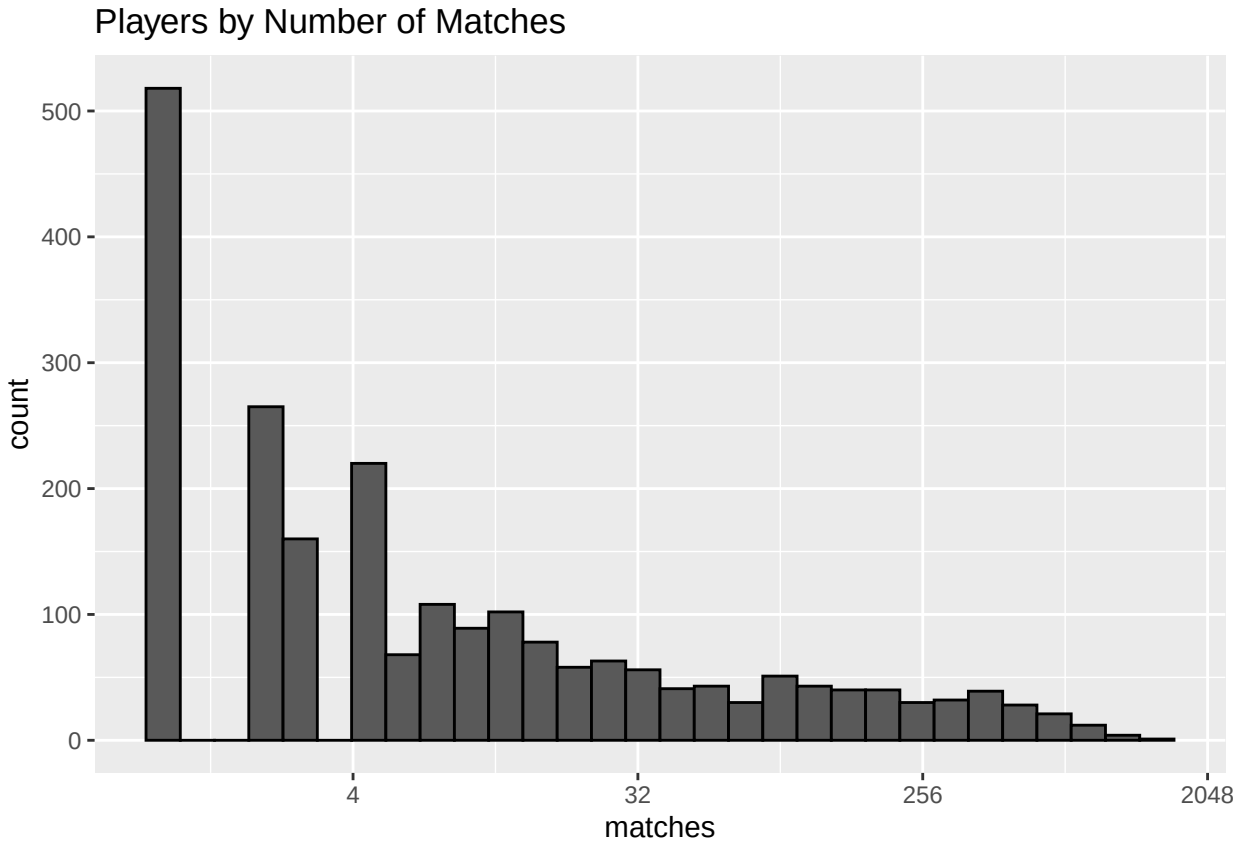
Introduction

The goal of this project is to predict the winners of tennis matches based on previous matches. A baseli

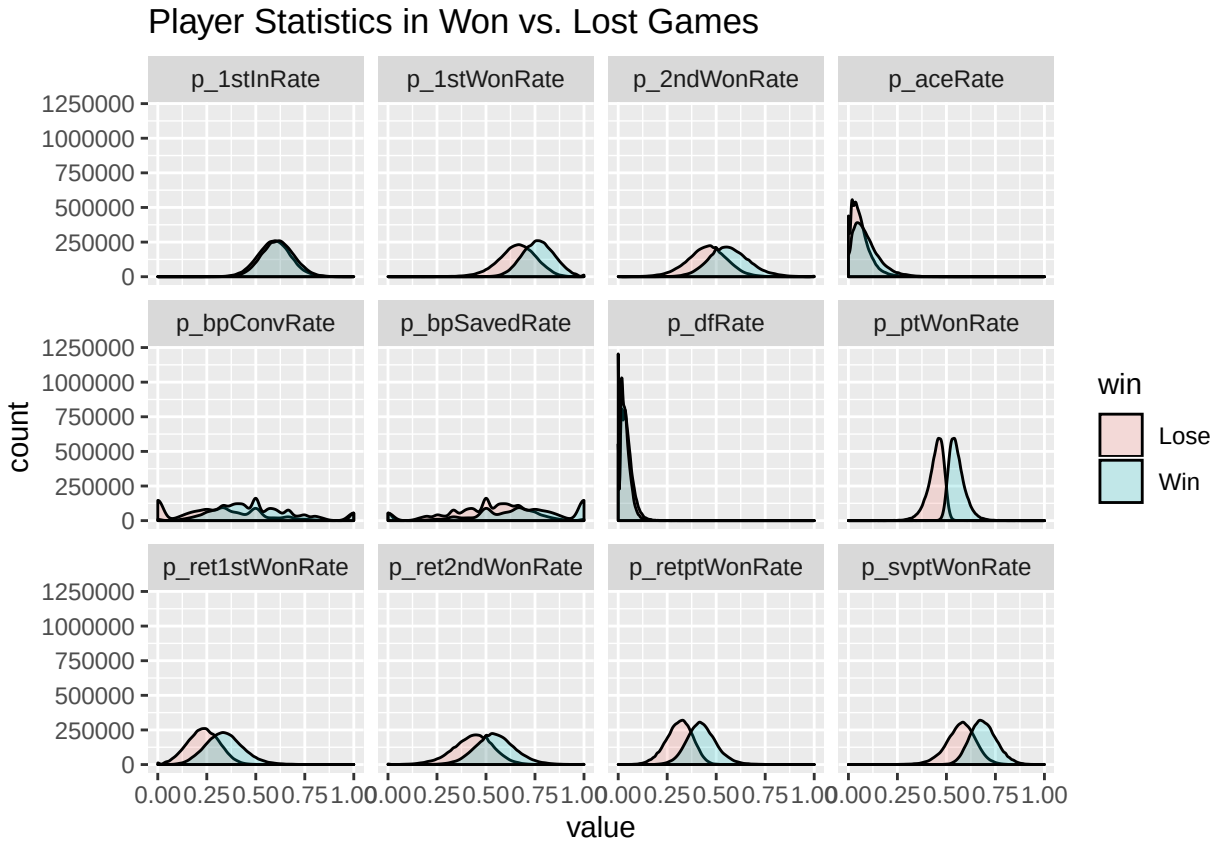
Methods

The base data required tidying before being usable for match prediction. Any matches that are walkovers
The base data also needed to be reformatted to be used for match prediction, as the entries are encoded
Some insights can be gained from examining the base data given, such as the fact that there are a consi





We can also examine some basic statistics and how they are related to winning matches. Looking at the plot, we can see that statistics like the amount that a player double faults and the amount of 1st serves that are in bounds have little bearing on the outcome of the match. Statistics like the percent of 1st serves won and 2nd serves won are somewhat related to the outcome of the match, and unsurprisingly the fraction of total points won is strongly connected with the winner of the game. Interestingly, there is some overlap in the fraction of total points won between won and lost games, as it is possible to win a match without getting a majority of the points. This happens in about 5% of the matches in the dataset.



A baseline prediction accuracy is established by counting player wins and losses, and calculating the percent of games won by each player. The winner of any given match is then predicted to be the player with the higher win rate. A second prediction model is also made by predicting based on player average fraction of total points won, which gives slightly different results as the two are not perfectly correlated. To compete with the simple win rate model, a model based on player base skill ranking is created. This is based on the intuition that the outcomes of matches between players of relatively similar skill are more informative than outcomes between players with large skill gaps. Player rankings are computed by using the Elo rating system while iterating over the matches in the training set. The K update constant used in the Elo rating system is tuned over the training set. In an attempt to improve on the basic rating system, a version of it is also tried where the actual outcome is given as the fraction of total points won, rather than simply assigning wins a value of 1. This is based on the intuition that players with close skill levels are expected to have close matches, whereas players with a large skill advantage are expected to win by a larger margin. The weight of the score versus who actually won is tuned for values between 0 (score is ignored) and 1 (only score is used). Some additional factors for examination are calculated by computing players' average statistics and then finding the difference between the two players. Players' offensive statistics are the difference between their serving skill and their opponents returning skill, and defensive statistics are the difference between the players returning skill and their opponent's serving skill. The difference in player skill rankings is also included. These factors are plugged into a generalized linear model to predict the winner.

Results

The accuracy of each model is compiled in the table below:

```
##                                    method  accuracy
## 1:                       Random Guessing 0.5172545
## 2:                     Highest Win Rate 0.6321578
## 3:           Highest Avg Point Won Rate 0.6283656
```

```
## 4:                                Elo Rating System 0.6549109
## 5: Elo Rating System w/ Weighted Points Won Rate 0.6575654
## 6:                                           GLM 0.6499810
```

The Elo rating system based approach only provided marginal improvements over simply using win rate as a

Conclusions

A few factors were tried to approximate player skill and improve on the baseline prediction accuracy, bu