

DATA 621—Business Analytics and Data Mining

Fall 2020—Group 2—Homework #1

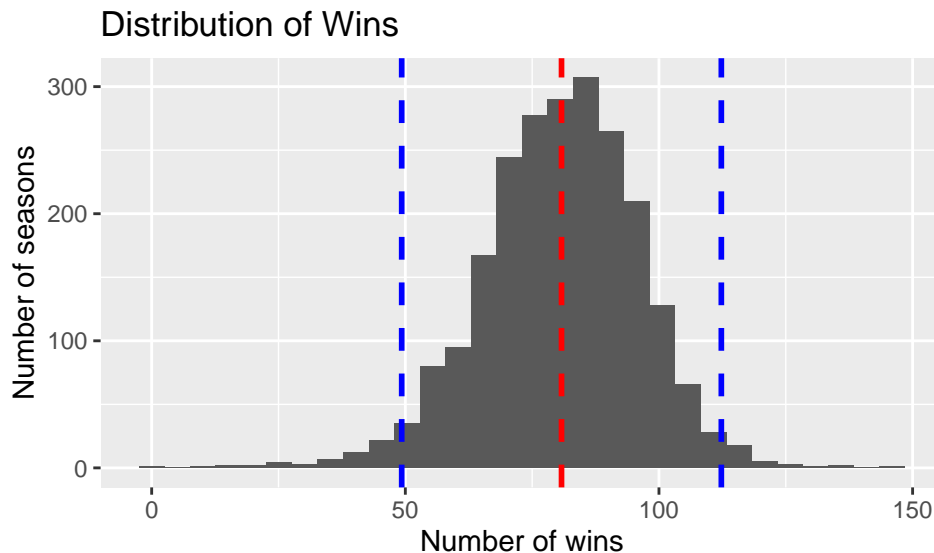
Avraham Adler, Samantha Deokinanan, Amber Ferger, John Kellogg, Bryan Persaud, Jeff Shamp

9/27/2020

1. DATA EXPLORATION

How often does the team win?

We are given a data set containing 2,276 records, each of which includes 15 statistics and the total number of wins a team had in a given year. On average, the team wins about 50% of the games they play (81 games out of 162), with the best season having 146 wins and the worst season having 0 wins. The data is normally distributed and most seasons have between 49 and 112 wins (blue lines below). This means that there aren't too many extreme seasons where the team won significantly more or significantly less games than usual. This serves as a good gut-check for our final predictions; if the predicted wins are too high or too low, we know something is probably off.



What's missing?

A first look at the data shows that only about 8% of the records have a full set of information. The good news is that most of the missing values come from statistics that don't happen too often: hit-by-pitch (92% missing!), caught stealing (34% missing), and double plays (13% missing). Since we have so little hit-by-pitch data, we expect that it doesn't contribute much to overall wins and will eliminate it from a few of the models we propose. The other two stats have less than half of the data missing, so we'll need to think of a clever way to fill in these values. The remaining missing information is from a combination of stolen bases and

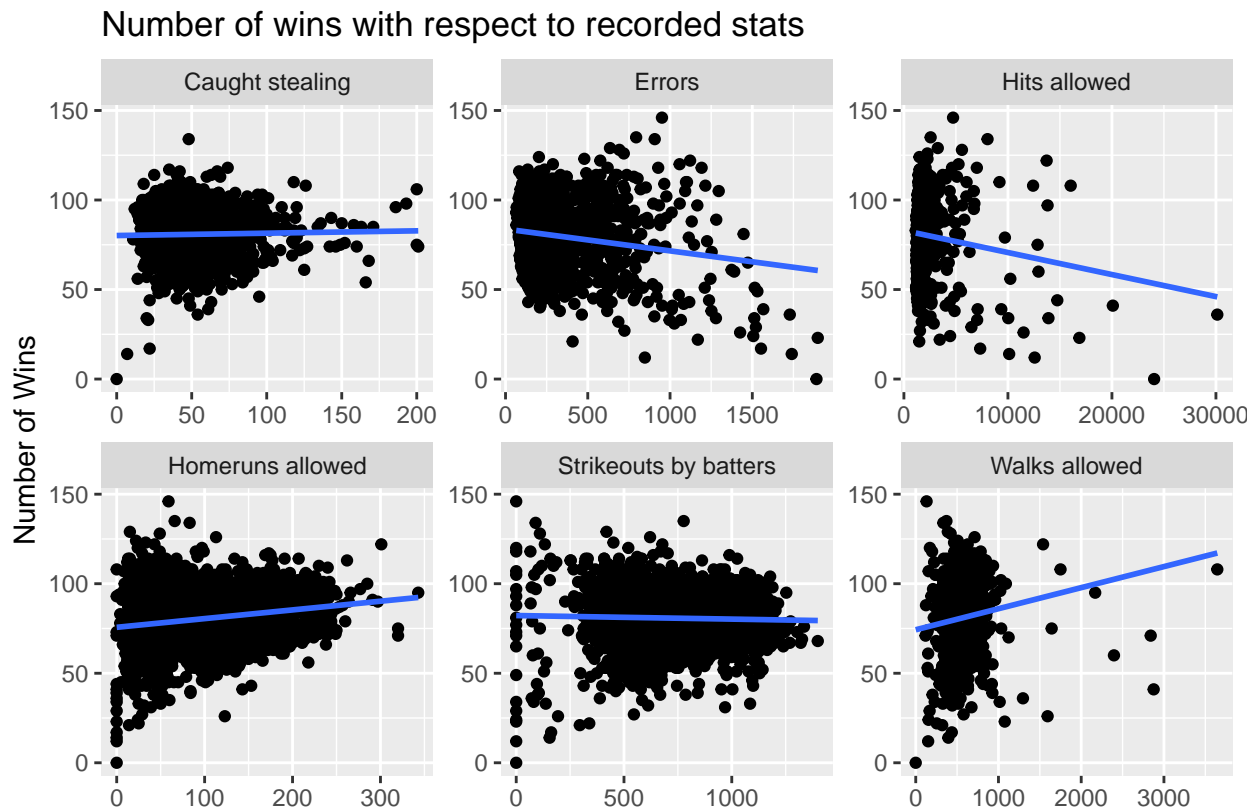
strikeouts (by batters and pitchers). **It seems completely unreasonable** to have zero strike outs in a season, so this is something we'll most certainly have to impute.

Do the individual stats affect winning?

Intuitively, we expect that Caught stealing, Errors, Hits allowed, Homeruns allowed, Strikeouts by batters, and Walks allowed would all have a **negative** impact on the total wins. In other words, as these values increase, we expect that the team is less likely to win.

However, when we take a closer look at the data, these relationships aren't obvious. In fact, only Errors and Hits allowed seem to have a negative impact on wins. Caught stealing and Strikeouts by batters appear to be random; this means that whether the stat for a particular season is high or low doesn't affect the overall number of wins.

Even more interestingly, Homeruns allowed and Walks allowed have the *opposite* effect; as these stats increase, so do the number of wins!

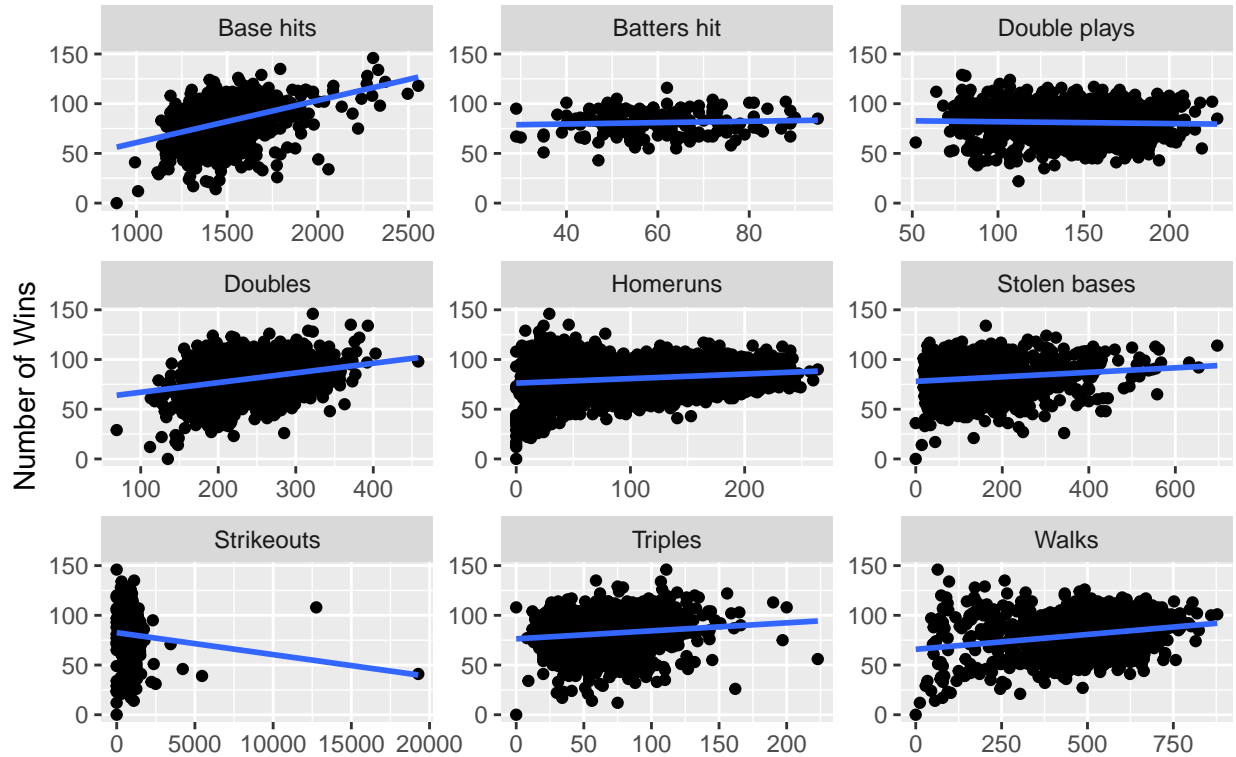


We can look at the same information for the stats that we expect to have a **positive** effect on wins: Base hits, Doubles, Triples, Homeruns, Walks, Batters getting hit by pitches, Stolen bases, Double Plays, and Strikeouts by pitchers.

Many of these stats *do* seem to have an effect on the number of wins. Most notably, as the number of base hits and walks increase, so do the number of wins. We see weaker positive relationships for homeruns, doubles, triples, and stolen bases. This makes sense when we think about it – these things tend to happen less often in games than pure base hits and walks, so they don't have as much of an effect on winning. Double plays and batters hit don't appear to have an effect on the number of wins either. Once again, this intuitively makes sense because they are less likely to happen in a game.

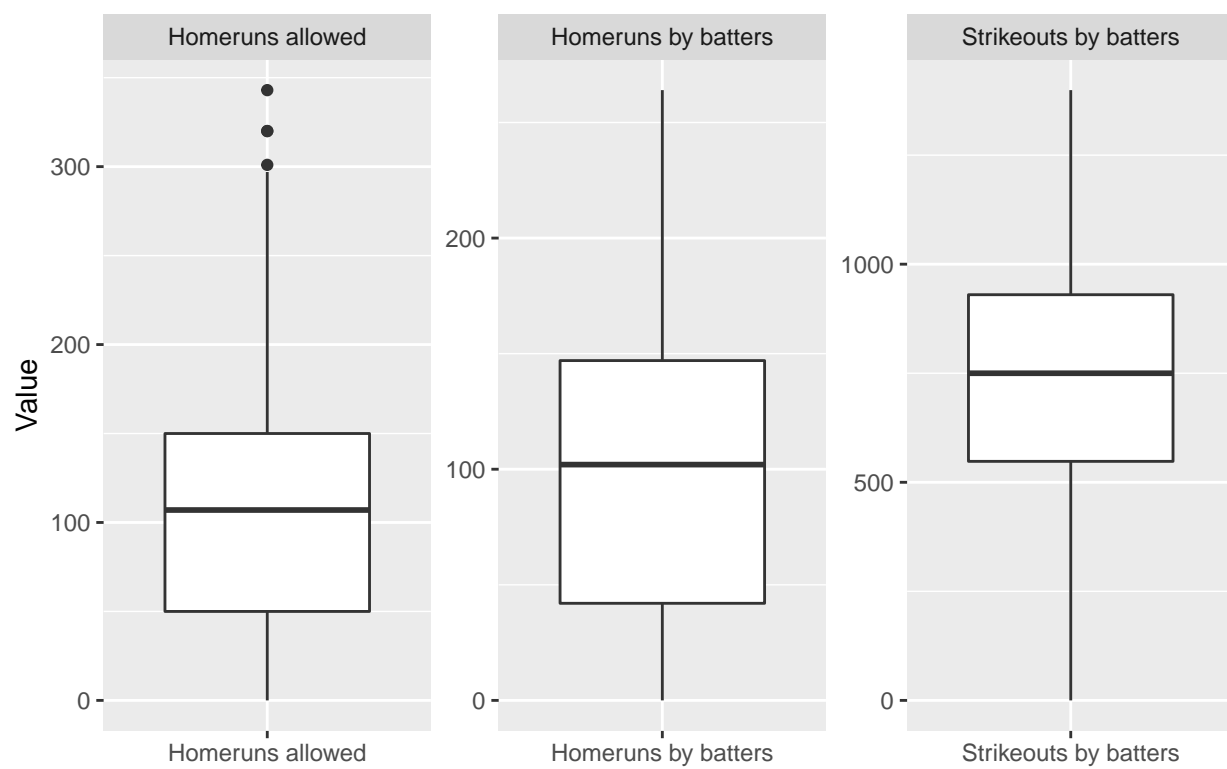
One thing to note is the number of strikeouts compared to the number of wins. We can see that there are a few outliers – abnormally high numbers of strikeouts in a season. This should be taken with caution, as they don't represent a typical season's stats.

Number of wins with respect to recorded stats



Are some stats more skewed than others?

Some of the provided statistics are well-balanced in the sense that there are very few (or no) extreme values. **Homeruns by batters**, **Strikeouts by batters**, and **Homeruns allowed** are examples of this. Aside from



Are stats correlated?

