DATA 621—Business Analytics and Data Mining

Fall 2020—Group 2—Homework #1

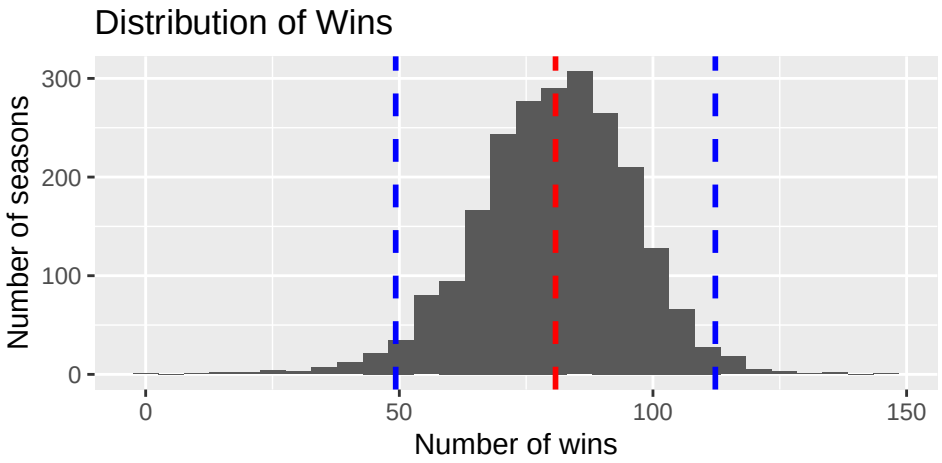
Avraham Adler, Samantha Deokinanan, Amber Ferger, John Kellogg, Bryan Persaud, Jeff Shamp

9/27/2020

1. DATA EXPLORATION

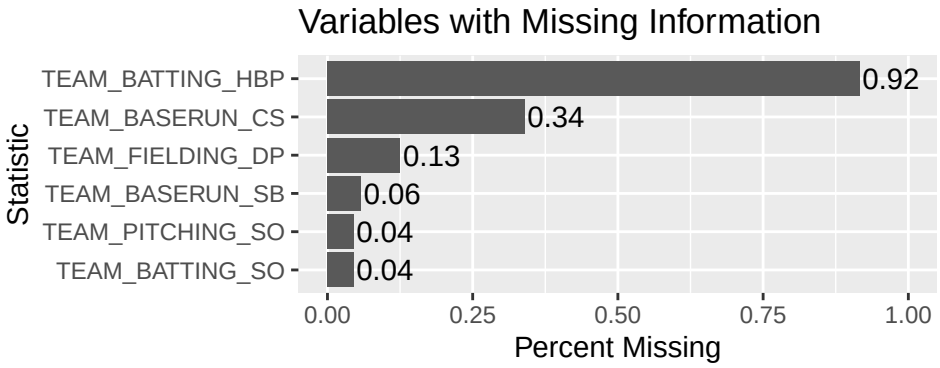
How often does the team win?

We are given a data set of 2,276 records containing 15 seasonal statistics and the total number of wins a team had in a given year. On average, about 50% of games played are won (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 years have between 49 and 112 wins (blue lines below). The nature of the distribution means there aren't too many extreme seasons where wins are significantly higher or lower 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 in our model is probably off.



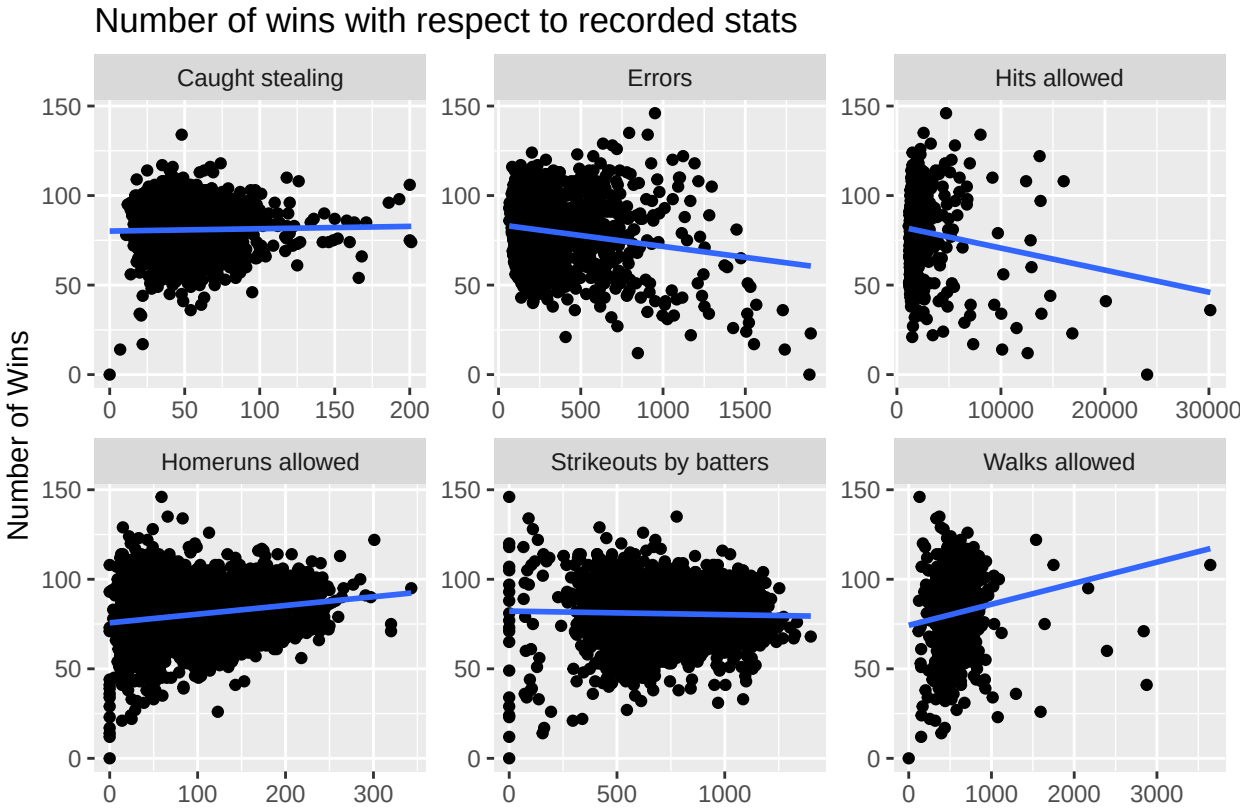
What's missing?

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?

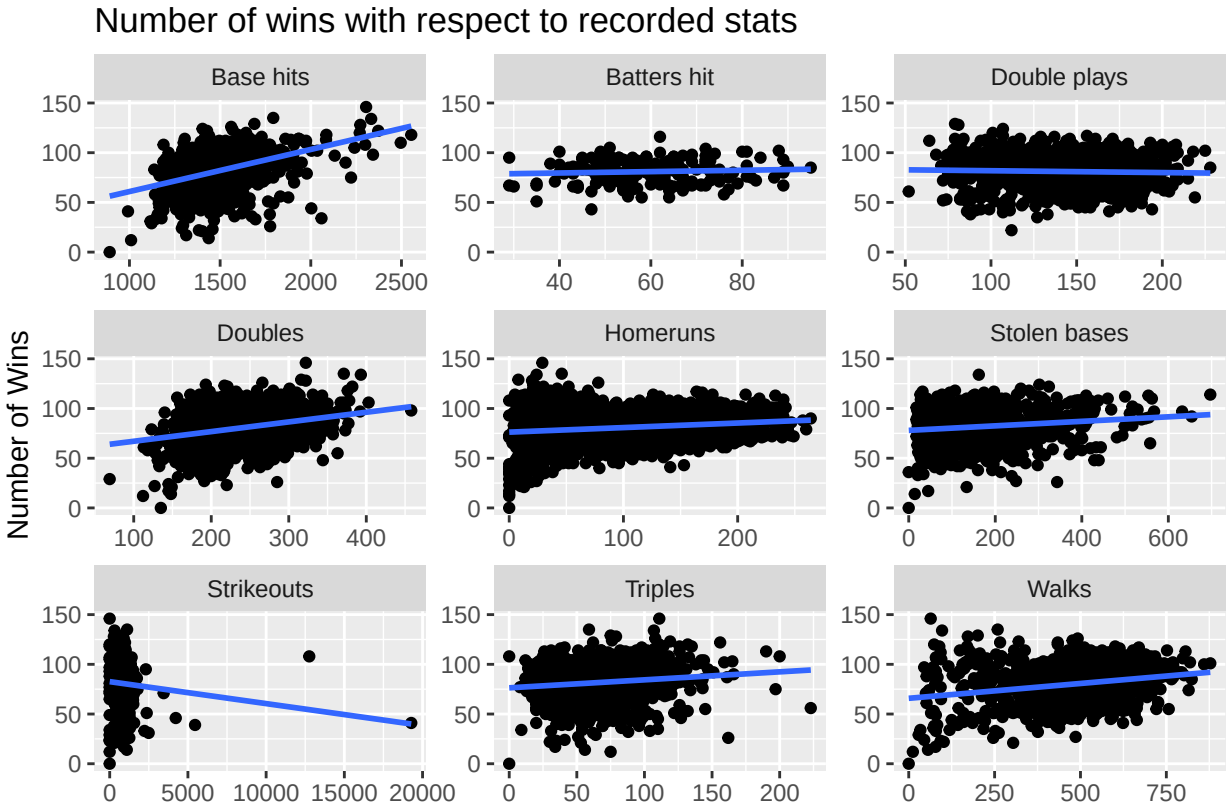
Stats with an expected negative impact: Intuitively, we expect that Caught stealing, Errors, Hits allowed, Homeruns allowed, Strikeouts by batters, and Walks allowed would all have a **negative** impact on the total wins. In other words, as these values increase, we expect that the team is less likely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



When we take a closer look at the data, these negative 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!

Stats with an expected positive impact: 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.



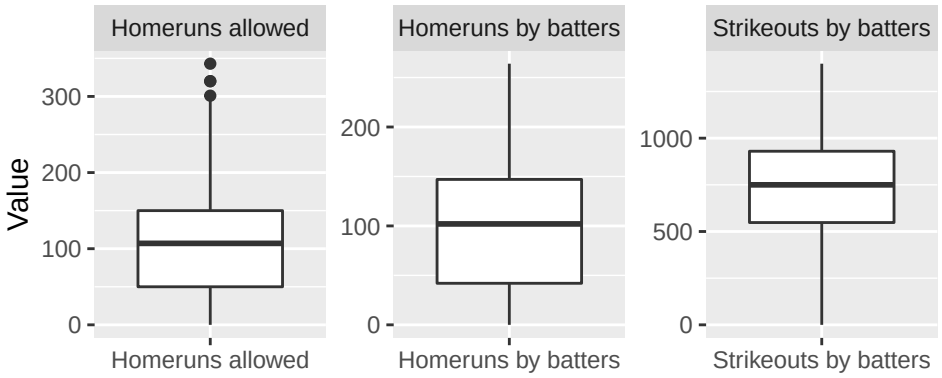
Many of these stats *do* seem to have an effect on the number of wins, most notably, base hits and walks. We see weaker positive relationships for homeruns, doubles, triples, and stolen bases. This makes sense when we think about it; these things tend to happen less often in games than pure base hits and walks, so they don't have as much of an effect on winning. Finally, double plays and batters hit don't appear to have any correlation with the number of wins. 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.

Are some stats more skewed than others?

Before using any of the statistics in a model, we need to take a closer look at the variation in the data. We call out of the ordinary values (exceptionally high or low values) **outliers**. We need to take these into account in our modeling because we want to make sure our predictions aren't skewed because of them.

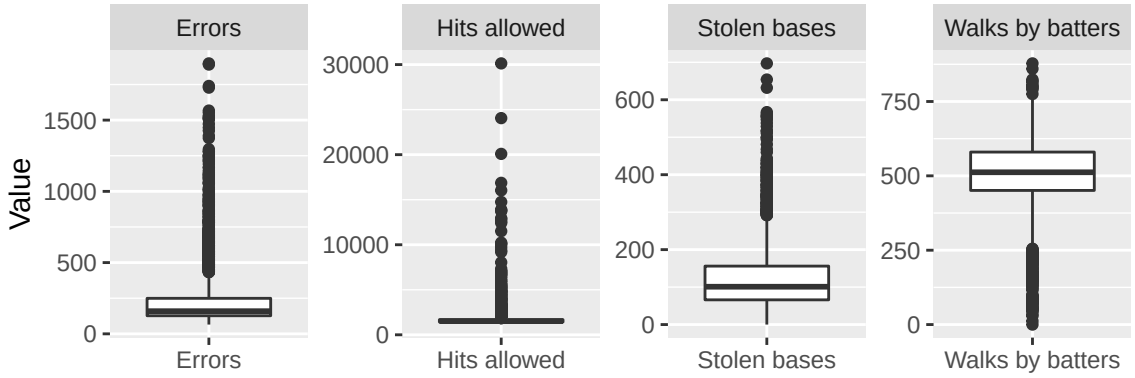
Some of the provided statistics are well-balanced in the sense that there are very few (or no) extreme values. **Homeruns by batters**, **Strikeouts by batters**, and **Homeruns allowed** are examples of this.



Some things to note about each of these statistics:

- **Homeruns allowed** (average = ~100/year) and **Homeruns by batters** (average = ~106/year) have a very similar mid-range distribution (50% of the data lies between ~50 and 150). The slight difference in average stats means that teams tend to have a higher number of homeruns than the opposition team.
- The only thing that stands out about **Strikeouts by batters** (average = ~736/year) is how nearly perfectly normal it is. 50% of the data is between about 500 and 1000 and there are absolutely no outliers in the dataset! This means that there were no surprisingly high or low seasons.

Conversely, some of the stats have a very high number of outliers, indicating that there are some seasons with some abnormally high or low values. **Errors**, **Hits allowed**, **Stolen bases**, and **Walks by batters** are examples of this.



Some things to note about each of these statistics:

- All of the outliers for **Errors** and **Hits allowed** are above the upper tail of the data set. This is further illustrated by the mean and median values for both of these stats; in both instances, the mean per year (Errors = ~246/year and Hits allowed = ~1779/year) are higher than the median per year (Errors = ~159/year and Hits allowed = ~1518/year). This means that some seasons with exceptionally high values for both of these statistics skew the dataset.
- The maxiumum value

Are stats correlated?