Marcin Tomala

CISC. 3225

Professor Levitan

22 September 2021

Ethical Audit

## Demographics

• Demographics. Are all groups fairly represented? Are inappropriate questions being asked? Are other variables being used as proxies for race?

The included dataset encompasses Major League Baseball (MLB) teams’ data over the period from 1962-2012. Each entry represents a team’s season, with its win-loss totals, playoff appearance (or lack thereof), final season ranking, and batting statistics. Since the entries in the dataset do not represent individuals but rather aggregated team statistics, there is no concern that the data portrays different demographics - explicitly or via proxies - in a way that would create an unfair representation. Moreover, even if the dataset did represent individual players and their statistics, there would still be no way of identifying each player’s demographic based on the columns, as the metrics are only concerned with performance on the field. The questions being asked are also unconcerned with the demographic makeup of the league or anything else related to demographics.

## Missing Data

• What is missing from your data? Look into how it was collected and the criteria for inclusion. Is data from some groups or neighborhoods more likely to have fields missing or to not be in the data at all? How does this affect those groups or neighborhoods?

The data was collected uniformly from teams’ appearances in MLB games across each season. Since the structure of the MLB fluctuated throughout the years, not every season’s data is comprised of entries for the same set of teams: some teams have relocated, others - founded and incorporated into the league, expanding the dataset, particularly as the years go on. Nevertheless, save for the seasons shortened by players’ strikes, there are no omissions in the dataset. While excluding a team or a few teams from the dataset would likely not have any effect on them and their performance in the real world, it could skew the analysis concerning the desired outcome - ascertaining which statistics are of the highest value for guiding efficient yet effective team building.

Even though the two columns that represent data for certain opponents’ statistics (Opponent’s On-Base Percentage and Opponent’s Slugging) have only been included for the years 1999-2012 (presumably only having been tracked since 1999) the fact that these data points are missing for the years 1962-1998 does not necessarily invalidate or skew the conclusions that can be drawn from the remaining columns for that period. The years which have these stats included might turn out to be more revealing regarding what it takes for a team to succeed in the MLB, however.

## Data Usage

• How is this data being used? What kinds of questions can this data be used to answer? Should they not be asked? How likely is it that they can be accurately answered? How likely is it that they will be used to hurt people?

This dataset can be used to answer questions regarding team-building in professional baseball. As documented in the book (and the movie) *Moneyball*, in the early 2000s, financial constraints that kept their team from being competitive in the league (in the face of significantly richer teams such as the New York Yankees or Los Angeles Dodgers) led the executives of the Oakland Athletics to look for other, lesser-known ways to evaluate players. Using datasets likely similar to this one, they have concluded that On-Base Percentage (OBP) and Slugging Percentage (SLG) were also reliable predictors of teams’ success, going against the conventional wisdom of the time that valued primarily batting average. With this information, they were able to sign individuals with high OBP and/or SLG whose batting averages were not particularly impressive, driving their price down. This approach to team-building led the A’s (as well as other MLB teams who adopted similar strategies later) to remain competitive (103-59 win-loss record in 2002) despite their low payroll - implying that there does seem to be some merit in this kind of analysis as a guide for the improvement of a baseball team.

There appear to be no immediate ethical concerns regarding these types of questions. Teams utilizing these metrics will not be gaining an unfair advantage over other teams, since the conclusions are drawn from publicly available data, allowing every team to follow suit in analyzing datasets like this one. On the contrary - these kinds of analyses can be seen as a leveling of the playing field for less affluent teams like the aforementioned Oakland Athletics. On a more granular level, players with high batting averages but comparatively less impressive OBP/SLG stats will not become more likely to lose their jobs when the focus shifts to other offensive metrics - after all, such players’ OBP and SLG should still be well within the line to keep them employed, as OBP is strongly correlated with batting average (given that the formula for OBP includes the ratio of hits to at-bats).

## Sourcing and Consent

• Where did you get this data? Privacy violations? Informed consent? Benefits (to subject or humanity) outweigh harms?

The dataset was shared publicly on [Kaggle](https://www.kaggle.com/wduckett/moneyball-mlb-stats-19622012) by a user who obtained it from [baseball-reference.com](https://baseball-reference.com). There are no apparent privacy violations in the collection of the data, given that it is sourced and aggregated from public events (baseball games) and only concerns franchises rather than individuals. The dataset and its analysis are unlikely to cause any harm to the subjects - historical data about team performance will not (and likely cannot) be used in a way that will affect teams or their employees in any meaningful way going forward. The statistical work done on datasets like this one is primarily concerned with studying new avenues to apportion a team’s resources on the path to competitive success. While the success of one team does inherently negatively affect the success of another’s (only one team can win), the data can only *assist* in team-building and cannot directly affect the outcome of games.