**Project explanation:**

**1. Initial problem statement**

First, to summarize my final project in a single sentence, I will be working through historical Major League Baseball statistical data to determine which statistics/benchmarks early in careers are predictors of future success.

The target audience consists of general managers looking to figure out which players they should invest in early on (whether that means trying to trade for them, or signing them to a contract before they hit the free agent market).  This is a particularly relevant issue since player salaries have risen for a very long time, but as teams have recently become much more analytically-minded, what they've found is that paying players in their 30s for what they did in their 20s is not a particularly profitable strategy.  The goals of this project, then, are both to determine which statistics are good predictors of success, and to provide recommendations of current players to target based on those findings.

As far as success, it can be measured in myriad ways, but I will be looking at it from the perspective of the statistics of those who went on to have careers statistically in the top x% of all MLB players (the statistic I will use to slice it is WAR-- Wins Above Replacement-- a new but widely-respected aggregate statistic that attempts to determine the number of wins a certain player is worth above a replacement-level player).  The idea of what characterizes "early on" in a player’s career is also variable, but I will be looking at it both from the perspectives of games played and x-number season (e.g. 3rd Major League season).

The primary data source I will be using is the Baseball Reference Play Index ([https://www.baseball-reference.com/play-index/](https://www.baseball-reference.com/play-index/" \t "_blank)), an extensive (and possibly overwhelming if you don't love this stuff like I do) collection of historical MLB statistical data from 1871 to the present.  It contains data ranging from single-season and career statistics (both basic and advanced) for batters, pitchers, and fielders, to event finders that allow you to pick out all the instances of a single play or set of plays in a game, in a season, or in the history of MLB.  The beauty of this source is that it allows you to make very specific queries to find the information you're looking for (e.g. if you wanted a list of all the instances that a player born in Alabama had a multi-homerun game between the years of 1940 and 1954, you could make that query and find that there were 6 different instances of that along with the stats from those games, then download the table as a CSV).

1. **Methodology summary: a)** looked at top 100 players by WAR for hitters and for pitchers, **b)** took averages of their first 3 year stats and compared it to those of the average player, **c)** used percentiles to determine which statistics were better predictors of success, **d)** recommended current young players whose corresponding statistics are highest
2. **Top 100 players: a) hitters-** after filtering out pitchers for hitter statistics (played at least 50% of games at non-pitcher positions), sorted by offensive WAR (excluding defense, since the statistics I measured were purely offensive ones)—found 4,328 records of careers starting 1969 or after (I chose 1969 because I wanted to slice data to be more relevant, and that was when the most significant rule change in recent history happened—the raising of the pitchers mound), and pulled out the top 100 (a bit more than the top 2%), **b) pitchers-** after filtering out position players who pitched a few innings in lopsided games by setting IP > 10, sorted by pitcher WAR—found 4,217 records of pitcher careers starting 1969 or after (a bit more than the top 2%)
3. **Pulled those top 100 players’ stats from their first 3 seasons—***note:* *only 46 of the 100 hitters had first-3 WARs in the Top 200, 69 in Top 400; 52 of the 100 pitchers in the Top 200, 66 in Top 400*
4. **The most significant roadblock occurred here: Taking descriptive data of the average baseball player proved impossible,** as Baseball-Reference’s database was not sophisticated enough to provide a dataset of every player’s first 3 seasons (from 1969)—each of their results would provide a page of 200 players, and it was impossible to flip past page 9 (player 1800). **My solution** was to treat players with an OPS+ and ERA+ of 100 (for hitters and pitchers, respectively) as the average player (two very important statistics, the first being on-base percentage plus slugging percentage, and the second being Earned Run Average over 9 innings which are, after accounting for external factors like ballparks played in, adjusted so that 100 is the league average (so 150 would be 50% better than the league average, etc.)). The inherent issue is that anything which factors into those stats in my analysis is less credible (e.g. if I were to say ERA is heavily correlated with success, while that is undeniably true, it would be a self-fulfilling prophecy, since I’m comparing the successful players to an average player with a distinctly average ERA). Thus, any analysis of ERA, OBP, SLG, and OPS had to be taken with a massive grain of salt. Still, **this solution is perfectly warranted for the purposes of this analysis because it is merely comparative** (i.e. if we measure the percentiles of the average “Top 100” player in their first 3 years against our presumed “average” player, we don’t have to say that the Top 100 are in, say, the 85th percentile of all players, a claim we can’t make due to the limitations of this data, but simply determine that “Statistic A” is more significant in what makes a Top 100 player than “Statistic B”). I then widened the sample to include players with values between 95 and 105, to create a larger sample to measure against.
5. **Rate stats and necessary calculations:** At first glance, two things were very apparent in the comparison between the “average player” early career statistics and those of the “Top 100”: **a)** the “Top 100” were better in every category (not a surprise—the best players were better in their first 3 seasons than the average player), **b)** the “Top 100” played in many more games and pitched many more innings, which would skew any statistic that wasn’t a **rate statistic**, making it necessary to use those exclusively, and to calculate some myself. Statistics I added in were K/9, BB/9, HR/9 (I had to calculate these myself in some cases, as BBREF had these stats, but only let you throw in a certain number of parameters that they didn’t automatically include), and WHIP for pitchers, and for hitters I had to add SB%, as well as calculating runs, HR, and RBI into per AB stats so I could use them more effectively in analysis. **I also calculated averages** (mostly just taking the average of the row, but I had to do much more complex calculations (SLG was a particularly rough one) for any stats that were already rate stats (e.g. you can’t just average up all the batting averages and say that’s the average batting average, because they were taken from different denominators—in that case you have to add up all the hits, and divide them by all the ABs).
6. **Tableau analysis:** Looked at where the “Top 100” averages for each statistics lay on the “average player” distribution, choosing 9 stats (all rate stats) for both pitchers and hitters, each getting its own dashboard: a) **Hitter stats-** BA, BABIP, SB%, R/AB, HR/AB, RBI/AB, OBP, SLG, OPS, b) **Pitcher stats-** Winning Percentage, HR/9, K/9, BB/9, K/BB, WHIP, FIP, BABIP, ERA
7. **Findings:** Unsurprisingly, the ERA and OPS of the “Top 100” had the highest percentiles across the “average” distributions (which, as mentioned, must be taken with a grain of salt)—**beyond those,** the most significant hitting statistics were **batting average** and **R/AB**, while for pitchers they were **WHIP** and **K/BB ratio** (FIP followed closely behind, but since that’s another adjusted ERA, I largely discounted that)
8. **Player recommendations:** Looked at hitters between 300 and 500 games played currently (since this would be a range where they have played a large enough sample of games where one might be able to determine how good they are, but small enough that they may not have signed a big contract yet), and pitchers between 50 and 85 games started (I excluded relievers, since they are largely not present in our original WAR analysis). I chose to do this, rather than seasons played, because it is actually a more scientific measure (one could start a season at the end and it would still count as their first season), and unlike with the larger datasets, Baseball Reference did allow me to slice active players in this way. These results yielded 93 hitters and 58 pitchers, after I filtered out anyone who was listed as active but hadn’t played since 2016. *In the final two tabs of the attached Excel spreadsheet*, I ranked the players by the pair of stats (for each) found to be most important. I also did not want to leave out OPS or ERA (since they are indeed determining factors of success), but I weighted them less, dividing their rankings by 2 (OPS/ERA Adjustment columns in the table). I then added these three rankings in each, and used it to rank the top pitchers and hitters **(found at the bottom of each worksheet)**. *Note: In the hitter worksheet, I also included Michael Conforto, the 12th ranked offensive player, because he was in the top 10 of OBP, SLG, and OPS.*
9. **Limitations: a)** the OPS+/ERA+ issue, which has been discussed in detail, but again is justified by the comparative nature of this analysis, **b)** relatedly, every single one of these statistics is used to derive WAR, so that is an inherent flaw in using them to predict it, **c)** the analysis excludes defense (since it’s significantly less important in determining a position player’s WAR, but it is still a determining component of a player’s value), **d)** the analysis largely excludes relievers, since they pitch less innings and therefore accumulate less WAR, **e)** I would have strongly preferred to look at a player’s first x games rather than their first x seasons because it’s more scientific (as mentioned previously), but the data would not allow me to slice it that way

**Data dictionary:**

**WAR** (Wins Above Replacement) *(used to determine Top 100 pitchers and hitters in this project)*- aggregate statistic that uses a number of factors to determine the number of wins a certain player is worth above a replacement-level player. Full explanations of pitcher WAR ([**https://www.baseball-reference.com/about/war\_explained\_pitch.shtml**](https://www.baseball-reference.com/about/war_explained_pitch.shtml)) and position-player WAR ([**https://www.baseball-reference.com/about/war\_explained\_position.shtml**](https://www.baseball-reference.com/about/war_explained_position.shtml)**)** offered here—*note: this project uses oWAR (offensive WAR), which excludes a position player’s defense, because defense was not part of my analysis*

**Rk** (Top 100 Pitchers/Hitters First 3 Years)- WAR ranking among all [pitchers/hitters] first 3 years (since 1969)

**Other pitcher statistics:**

**IP-** Innings Pitched

**G-** Games

**GS-** Games Started

**W-** Wins

**L-** Losses

**W-L%-** Winning Percentage

**H-** Hits Allowed

**ER-** Earned Runs Allowed

**ERA-** Earned Run Average per 9 Innings

**ERA+-** Adjusted ERA (accounting for external factors such as ballparks played in) so that 100 is the league average (so 150 would be 50% better than the league average)

**SO-** Strikeouts

**K/9-** Strikeouts per 9 Innings

**BB-** Walks Allowed

**BB/9-** Walks Allowed per 9 Innings

**K/BB-** Strikeout to Walk Ratio

**WHIP-** Walks + Hits per Inning Pitched

**FIP-** Fielding Independent Pitching (ERA adjusted to be independent of defense, i.e. factors out of the pitcher’s control)-- (13\*HR+3\*(HBP+BB)-2\*K)/IP

**BAbip-** Batting Average Allowed on Balls in Play

**HR-** Homeruns Allowed

**HR/9-** Homeruns Allowed per 9 Innings

**BF-** Number of Batters Faced

**Other hitter statistics:**

**G-** Games

**PA-** Plate Appearances (includes walks and sacrifices)

**AB-** At-Bats (does not include walks and sacrifices)

**R-** Runs

**H-** Hits

**2B-** Doubles

**3B-** Triples

**HR-** Homeruns

**RBI-** Runs Batted In

**BB-** Walks

**SO-** Strikeouts

**BA-** Batting Average (Hits per AB)

**OBP-** On-Base Percentage (Percent of Plate Appearances On Base) (Hits + Walks + Hit by Pitch) / (At Bats + Walks + Hit by Pitch + Sacrifice Flies)

**SLG-** Slugging Percentage (Total Bases per AB)

**OPS-** OBP + SLG

**OPS+-** Adjusted OPS (accounting for external factors such as ballparks played in) so that 100 is the league average (so 150 would be 50% better than the league average)

**BAbip-** Batting Average on Balls in Play

**HBP** (included in some because part of OBP calculation)**-**

**SF** (included in some because part of OBP calculation)**-**

**SB-** Stolen Bases

**CS-** Caught Stealing

**SB%-** Percentage of Successful Stolen Bases

**R/AB-** Runs per AB

**HR/AB-** Homeruns per AB

**RBI/AB-** RBI per AB

**OPS** and **ERA adjustment** (Current Hitters/Pitchers)- OPS/ERA ranks divided by 2 (calculated so they could be included in final ranking, but to a 50% lesser extent since they are a crucial part of the OPS+ and ERA+ calculations used to select the “average” player)