**🧠 Can We Predict MLB Player Value?**

I’ve always been curious of how professional sports teams decide what a player is worth — especially in baseball, where the numbers behind the game run deep. Think *Moneyball*, but pushed further into the modern age of data science and machine learning. What kinds of players get rewarded the most in today’s MLB? Can we use statistics to reverse-engineer how front offices assign value?

Using a combination of statistical modeling and machine learning, I set out to explore what drives player valuation in Major League Baseball. My goal is to build models that can predict a player’s contract value and length, while also using those predictions to assess whether players are overvalued, undervalued, or priced just right. To conclude the study, I’ll try my hand at projecting contracts for some upcoming big-name free agents, to see what it might cost to sign them right now. It is worth noting that because this is just an experimental project, it was completed with some limitations and assumptions, which I’ll mention as they come up.

Along the way, I’ll walk through:

* How I constructed a dataset by merging contract details with advanced player stats and career accolades, as well as personal variables like Instagram following.
* The exploratory analysis that helped uncover early insights and shape modeling decisions.
* K-means clustering to group players into prototypes — and see how each group tends to get paid.
* A correlation study to identify which features (like WAR, age, or hard-hit rate) have the most predictive power.
* Predictive models trained on pre-2025 contracts to evaluate the newest free agent class.
* Attempting to project what it would cost right now to sign some of the best players heading into free agency.

In general, the research followed 8 steps:

1. Download/Export hitter statistics from [**Baseball Savant**](baseballsavant.com).
2. Scrape contract data from [**Spotrac**](https://www.spotrac.com/mlb/contracts/).
3. Clean and merge contract data with statistics data.
4. Scrape **player accolades** from [Baseball-Reference](baseball-reference.com) and **fWAR** from [FanGraphs](fangraphs.com). Merge Instagram followers.
5. Perform exploratory analysis with visualizations.
6. Identify **player prototypes** with K-means clustering.
7. Conduct correlation analysis for feature selection.
8. Train and test **linear models**, assess player valuations.
9. **Project contracts** for upcoming free agents.

Whether you’re a baseball fan, a data scientist, or just curious about how teams make million-dollar decisions, this project has something for you.

**Let’s dive in.**

**🧱 Building the Dataset: Merging Performance with Paychecks**

In my opinion, the most time-consuming part of modeling is preparing the data. To model and predict MLB player valuation, I needed to create a dataset that blended performance metrics with contract outcomes. [Step 1] My starting point was player performance data from **Baseball Savant**, where I collected advanced hitting statistics for all position players who logged at least **50 plate appearances in a season between 2015 and 2024**. This gave me a wide base of player-seasons enriched with modern stats like xwOBA, barrel rate, sprint speed, and hard-hit percentage. I ran into my first limitation here – I could only get advanced hitting statistics from 2015 and beyond, as this was the year Statcast was officially implemented in all 30 MLB stadiums.

A screenshot of a computer

AI-generated content may be incorrect.

Figure 1. Preview Baseball Savant data.

[Step 2] Next, I scraped **the top 100 position player contracts signed in Major League Baseball**, using data from [Spotrac](https://www.spotrac.com/mlb/contracts/). This included details like total contract value, average annual value (AAV), contract start year, and team. I ran into my second limitation here – I had to settle for the top 100 contracts as Spotrac requires paid subscriptions for access to more data.

A screenshot of a computer

AI-generated content may be incorrect.

Figure 2. Preview Spotrac Top 100 Contracts.

[Step 3] To link performance data to contract outcomes, I:

* Assigned each player a “contract year,” defined as the season *before* their contract was signed — ensuring my models would never use data that came *after* a player got paid.
* Aggregated career metrics (totals, means, and rate stats) from the Baseball Savant data up to each contract year.
* Calculated improvements from rookie year to contract year to measure trajectory.
* Merged these stats with contract data using the player’s name and contract year as keys.

This process yielded a clean dataset of **91 out of the top 100 contracts**, where I had both performance metrics and contract details available. This is not a great amount of data for modeling, but it will have to suffice. There was some manual manipulation that I ended up performing – Vladimir Guerrero Jr. signed a 1-year arbitration deal after the 2024 season, and as of writing this, has now signed his mega-extension contract worth $500M over 14 years. In this case, I removed the record for his arbitration deal, and kept his extension, but changed the start of his contract to be 2024 rather than 2025.

[Step 4] With my new merged dataset of 91 players with contract outcomes and advanced hitting stats, I moved on to pulling WAR numbers, career accolades, and Instagram followers. I was able to do this with scraping methods and some manual labor, using three different sources:

1. [Baseball-Reference](https://d.docs.live.net/ccd545423950d79d/Documents/MLB_Player_Valuation/baseball-reference.com) for accolades like MVPs, World Series Championships, and Silver Slugger awards.
2. [FanGraphs](fangraphs.com) for WAR.
3. Instagram for follower count, to try to quantify marketability -- some players are going to provide a return to a franchise in the form of jersey sales, or higher attendance rates.

A screenshot of a computer

AI-generated content may be incorrect.

Figure 3. Preview Baseball-Reference accolades in the top right of the window.

I faced several obstacles here and had to settle for less-than-ideal data. For the career accolades, I struggled with scraping accolades *up to* the contract year. In the cases where the year of the achievement was present, I considered the award only if it was won on or before the contract year. If the year was not present, I counted the award. On [Baseball-Reference](https://d.docs.live.net/ccd545423950d79d/Documents/MLB_Player_Valuation/baseball-reference.com), if the player won an award more than once, you’ll know because the count followed by an “x” precedes the award text (see Figure 3 – “3x All-Star”, “2x Silver Slugger”). This is how I was able to tally how many of each award a player has won. The second obstacle was two-fold (not really an obstacle, but no easy solutions came to mind other than doing it manually) because I couldn’t get historic counts of Instagram followers and I had to collect the follower counts by looking up each player individually on my phone. As a result, I’m not able to quantify how many Instagram followers a player had before they signed. Do you think a player’s following goes up more pre-contract or post-contract? If the latter, the potential predictive power of Instagram followers could be over-estimated, and vice versa. On a good note, I was at least able to scrape WAR per player, per season. My data now has 109 variables to help predict player contracts.



Figure 4. Preview career accolades and marketability (top 5 contracts).

**🔍 Exploratory Analysis: Understanding the Player Pool**

[Step 5] Before jumping into modeling, I spent time exploring the data to better understand the types of hitters who land the biggest MLB contracts — and what sets them apart statistically. To keep this first pass readable and interpretable, I focused on each player's **career statistics**. The exploration certainly doesn’t cover everything, as this step can usually be limitless. I chose to look into things like contract trends over time, contract value vs age and advanced stats like WAR. I also wanted to see what kind of impact off-the-field performance has in terms of marketability and “putting butts in the seats”.

**📈 Summary Statistics: What Does a "Big Deal" Player Look Like?**

Using a basic statistical summary, I calculated the **mean**, **standard deviation (SD)**, and **coefficient of variation (CV)** for each feature. The CV — which measures the ratio of SD to mean — is especially helpful for spotting which variables vary the most relative to their scale. The table below shows the average value for each metric (mean), the average distance away, from the average, each value is (sd), and the percent variance of the standard deviation to the mean (cv).



Figure 5. Stat summary.

Here's what stood out:

* **Stolen bases** had the highest variance among all stats, suggesting a wide spread in baserunning ability among top earners.
* **HR/AB** (CV ≈ 35.2%) and **barrel rate** (CV ≈ 41%) also showed high variation — reinforcing that slugging is a major, but not uniform, driver of contract size.
* On the other hand, **sprint speed** and **exit velocity** were much more stable. Most players signing large deals ran between ~26.5 and ~29 ft/s and averaged 88–92 mph in exit velocity.
* On average, a player doesn’t sign a big deal until their age 27 season (5+ years in the big leagues), which means a player has to consistently perform through their arbitration years.

**💰 Trends Over Time: Are Contracts Shrinking?**

* The **average contract value** peaked in 2019 (Mike Trout and Bryce Harper signed new deals) at ~$230M and has steadily declined since. Even with Juan Soto’s massive $765M contract signed this year, 2025 had the lowest average value ($123M w/ Soto, $83M without) since 2018. Because there have been so few contracts signed, one of those being Vlad Jr. and his 14-year, $500M extension, as of now we’re seeing a pretty large increase in value starting in 2026.
* **Contract lengths** followed a similar drop — from 9.6 years in 2019 to just 4.8 years in 2025, but back up to 8 years in 2026. Again, among those in 2026 is a 14-year deal (Vlad Jr.) and a 9-year deal (Jackson Merrill).
* **AAV has remained relatively flat** since 2021, with a slight decline. The average is a bit skewed in 2020, as there were only two new contracts starting that year; Anthony Rendon signed for 7 years and $245M ($35M/year), and Christian Yelich signed for 7 years and $188.5M ($27M/year).

A graph of a line graph

AI-generated content may be incorrect.

Figure 6. Contract trends over time.

This suggests a shift in how teams are structuring deals: fewer mega-length commitments, but similar yearly payouts. Whether this reflects changing team strategy, collective bargaining shifts, or aging curves is open for debate — but the pattern is clear.

👴 **Age vs. Deal Structure: Youth Commands Longevity**

* There's a clear **negative correlation** between player age and contract length — the younger a player is at the time of signing; the more years teams are willing to commit. Younger players may carry more long-term upside, while older players command premium salaries in shorter time windows — reducing risk for the team.
* The positive trend in the left panel suggests that teams are willing to pay higher annual salaries to older players, likely due to proven performance — even if those deals are shorter.
* There are a few very high AAVs for older players and long contracts for some under-25 players, indicating special cases (e.g., MVP-level talents or early-career extensions).

A graph of a person playing baseball

AI-generated content may be incorrect.

Figure 7. Age x Contract Terms.

This dual-plot highlights how MLB teams structure contracts differently based on age. Younger players secure longer commitments, while veterans leverage their résumé into shorter, high-paying deals. From a modeling perspective, we can expect player age to be a significant predictor to contact terms.

🪖 **Declaring WAR**

* As we might expect, there's a clear upward correlation — players with higher career average WAR tend to earn larger total contracts.
* Some outliers exist in the top 4 overall contracts; **Shohei Ohtani** and **Juan Soto** stand far above the trendline, likely due to age, marketability, and projections of future performance — not just past WAR. The graph is named after **Mike Trout** for obvious reasons, he anchors the right side: high WAR, high contract value — essentially the gold standard. **Vladimir Guerrero Jr.** shows that even with modest average WAR, elite potential (and youth) can land a huge contract.
* **Underpaid overperformers likely exist** in the bottom-right quadrant — players with strong WAR but relatively lower contracts. The furthest-right dot in the graph that sits below the trend line is **Julio Rodriguez (5.79 WAR)**, who signed a 12-year extension worth over $209M starting in 2023 with the Seattle Mariners.

A graph with dots and lines

AI-generated content may be incorrect.

Figure 8. Mike Trout and then everyone else.

**📷 Quantifying Marketability**

As mentioned, I collected Instagram followers per player as of April 2025, so this should be taken with a grain of salt if you’re of the opinion that a player’s following will go up more *after* they sign their big deal. I think I find myself on the other side – the belief that a player’s popularity peaks right before they sign. While WAR and performance are core drivers, these charts suggest that public profile — potentially linked to marketability and endorsements — might add value at the negotiation table for superstar-level players.

A graph of followers

AI-generated content may be incorrect.

Figure 9. Shohei Ohtani is a global star.

* There’s a **general upward trend** — players with more Instagram followers tend to sign higher-value contracts. However, the relationship isn't perfectly linear, and there's considerable variance.
* Looking at the Top 4: **Shohei Ohtani** and **Juan Soto** are extreme outliers again, with both massive social media followings and record-setting contract values. **Vladimir Guerrero Jr.** and **Mike Trout** also stand out for having high contract values and solid followings, reinforcing the trend.
* On the other hand, **Starling Marte** has ~2.2 million followers on Instagram (around the same as Mike Trout) but is on the low end in terms of contract value.
* The bulk of MLB players cluster under 100K–500K followers, showing that massive social media presence is still relatively rare across the league.

A graph with black dots and blue lines

AI-generated content may be incorrect.

Figure 10. Instagram followers x contract value.

**🎏 K-means: Identifying the Player Prototypes that Demand Lucrative Deals**

[Step 6] Using machine learning methods like k-means clustering, I wanted to see if there were any patterns in player statistics and which player groups (prototypes) are likely to land larger contracts. K-means clustering is an unsupervised machine learning algorithm used to group similar data points into 'K' distinct clusters. It tries to partition the data such that the intra-cluster variation (within-cluster sum of squares, or WCSS) is minimized. One way to think of it is we’re trying to find 'k' centers (centroids) in the data so that each player is assigned to the nearest center, with the average distance from the points to their center being as minimized as possible, based on patterns in hitting statistics. This type of analysis can also be used to get insights into player valuations if we group and summarize them by average contract value. Will there be some players in some clusters making way more or less than the average for that group?

**🧮 Normalization: Leveling the Field for Modeling**

Before clustering, I normalized all numerical features using **z-score standardization**, a common technique that sets each variable’s mean to 0 and standard deviation to 1. This ensures that features with naturally higher numeric values (like home runs or career hits) don’t accidentally outweigh metrics like xwOBA or wOBA in models that rely on scale (e.g., linear regression or clustering). See snapshots below of data before and after normalizing.



Figure 11. Data before normalization.



Figure 12. Data after normalization.

**💪Throwing Elbows**

In order to cluster our players into groups, we need to mathematically try to figure out the optimal number of clusters (k). A popular and efficient technique to do this is by using what’s known as the ‘elbow method’, which will store the WCSS for 'k' clusters, up to 10 in my case. The lower the WCSS, the tighter the cluster. When looking at the plot below, the optimal number of clusters can usually be observed at the point in which the variance in the WCSS starts to shrink, where the line graph shows a distinct obtuse angle almost, hence ‘elbow’ – it’s difficult to tell here, but we can use the table on the right to see at which cluster center the variance starts to fall off. I went with 5, but you could make an argument for 4 as well. The difference clearly shrinks after 5, though, so we wouldn’t want to choose anything greater than that.

A graph with a line

AI-generated content may be incorrect.A screenshot of a table

AI-generated content may be incorrect.

Figure 13. Elbow plot for optimal k clusters.

**🏷️ Cluster Labeling and Descriptions**

Once we have the results, in the form of a cluster assignment for each player, we can start to interpret the player prototypes in each group by analyzing the statistical summaries.

Figure 14. Summary statistics for each cluster.

* **Cluster 1: All-Star Veterans**
  + Long contracts (~7.4 years) with high AAV ($26.9M)
  + Older age at signing (avg ~29)
  + Moderate HR output with high WAR (4.2 career avg)
  + Appear to be reliable, proven stars with accolades
  + Mookie Betts, Manny Machado, Francisco Lindor
* **Cluster 2: Franchise Superstars**
  + Longest contracts (~10 years), highest total value and AAV ($35M)
  + Young (~26), very high HR rate (0.08 HR/AB)
  + Highest WAR and All-Star/MVP recognition
  + These are elite, multi-tool players with big-market appeal
  + Juan Soto, Shohei Ohtani, Aaron Judge, Mike Trout
* **Cluster 3: Young with Upside**
  + Shorter, cheaper contracts (~$67M total), signed young (25)
  + Lower WAR and HRs, little to no accolades
  + Likely early-career players with room to grow
  + Ezequiel Tovar, Keibert Ruiz, Ceddanne Rafaela, Ozzie Albies
* **Cluster 4: High-Floor Contributors**
  + Medium-length contracts (~6.7 years), moderate value ($120M)
  + Solid production across the board
  + Younger, but still decorated, immediate All-Stars
  + Bryce Harper, Julio Rodriguez, Corey Seager, Matt Olson
* **Cluster 5: Power-First Sluggers**
  + Short-to-mid contracts (~5.7 years), solid AAV ($20M)
  + Older at signing (avg ~28.6), **high HR rate** (0.06 HR/AB), but high K%
  + Moderate WAR with some postseason/accolade history
  + Kyle Schwarber, Cal Raleigh, Jorge Soler, Joc Pederson, Rafael Devers

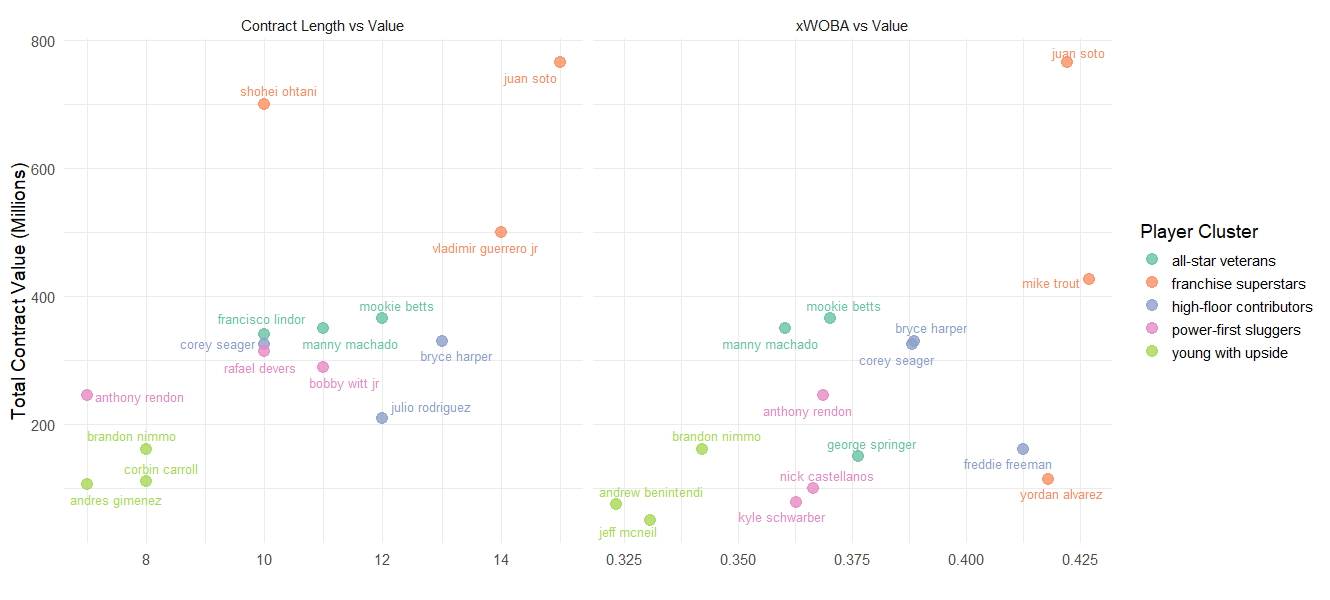
Now that I’ve interpreted the averages for each cluster and assigned cluster labels and descriptions, I want to move on to visualizing the clusters and look into outliers in each group.

Figure 15. Length x Value versus xWOBA x Value

* **Yordan Alvarez** (signed 6yr/$115M contract with Astros at age 25) seems to be really undervalued. It could be because he’s a true DH (Designated Hitter) and so does not add any defensive value to the franchise. However, his hitting metrics and career accolades are up there with the best in the game -- 0.0787 HR/AB, 3.4 WAR, 3x All-Star, Championship Series MVP, World Series MVP.
* **Freddie Freeman** (signed 6yr/$162M contract with Dodgers at age 31) also looks quite undervalued at a glance, but he was 31 when he signed, which probably played a factor. It is surprising not to see him grouped into the ‘All-Star Veterans’ cluster as an 8x All-Star with 1 MVP award and 3 Silver Slugger awards. He’s also a 2x World Series Champion and won a WS MVP.
* Another interesting observation is **Bryce Harper** and **Corey Seager** with nearly the same career xWOBA and contract values. Harper signed for 13 years and $330M with the Phillies when he was 25, while Seager signed for 10 years and $325M with the Rangers at age 27. Seager has the higher AAV and shorter contract length, maybe because he was 2 years older than Harper when he signed.
* **Brandon Nimmo** probably doesn’t belong in the ‘Young with Upside’ group – he was 29 when he signed his 8yr/$162M contract with the Mets. However maybe this group also includes [what I would consider] scrappy hitters, as the cluster on average owns a low K% and below-average power numbers. In the chart we can see their xWOBA is also on the lower end.

**📈 Correlation: Selecting the Features with Predictive Power**

[Step 7] Another limitation I have now is that my data only has 91 observations, but over 100 variables to potentially predict with. When features exceed the number of observations, it can lead to overfitting and poor generalization on unseen data. Calculating the correlation between features and my target variables will help highlight the features that have a strong linear relationship with my targets. In this case, my target variables are the contract value, and the contract length. Correlation values can range from -1 to 1, where -1 suggests a perfect negative linear relationship (as one goes up, the other goes down), and 1 suggests a perfect positive linear relationship (as one goes up, so does the other). A correlation value of 0 suggests no linear relationship.

**🧐 Pearson Correlation**

Pearson correlation measures the strength and direction of a linear relationship between two continuous variables. With some simple code in R, we can calculate the correlation of each of our features with the two target variables (value and length).

A graph of a graph with a bar chart

AI-generated content may be incorrect.

Figure 16. Correlated features with contract value.

* **Instagram followers** has the **highest correlation** with contract value (~0.55+), suggesting strong marketability and fan engagement may influence large contracts — even beyond on-field performance.
* **Career WAR** and **career xWOBA** rank next, showing that long-term performance and advanced offensive metrics are heavily rewarded.
* **Silver Slugger and MVP awards,** and **All-Star** selections are also positively correlated — postseason awards help validate perceived player value.
* Features like xWOBA, wOBA, xOBP, xSLG (both current and career) consistently appear, affirming that **underlying hitting quality** is closely tied to contract size.
* Traditional stats like average HR per season and career xSLG also show moderate positive correlation — **power still pays**.
* The **Catcher position** shows a slight negative correlation, suggesting catchers may earn less on average — possibly due to shorter peak windows or higher injury risk.
* **Player age** has a mild negative correlation, reflecting that **younger players tend to secure larger contracts** (likely tied to team control, longevity, and growth potential).
* Increases in at-bats and games played are slightly negatively correlated — this aligns with the idea that **wear-and-tear metrics** may reduce perceived future value.

A graph of a bar graph

AI-generated content may be incorrect.

Figure 17. Correlated features with contract length.

* **Career WAR** (0.54) is the strongest predictor of contract length — consistent high-level performance over a career is highly rewarded with long-term deals.
* **Other career batting stats** like wOBA, xWOBA, xBA, xOBP, and xSLG all correlate positively — clubs invest long-term in players with sustained offensive success.
* **Silver Slugger** (0.49) and **All-Star** (0.43) accolades also boost contract length — again reinforcing that recognized excellence contributes to longevity.
* **Player age** (-0.46) is strongly negatively correlated — older players tend to get shorter deals, which aligns with team risk aversion.
* **Seasons played** (-0.24) and **increases in at-bats** (-0.19) suggest that more wear-and-tear may reduce contract length — again, showing that durability concerns shorten deals.
* **Certain positions like DH** (-0.17) and **LF** (-0.18) are slightly negatively correlated — potentially because these positions are easier to replace or seen as lower-value defensively. This could help explain the questions I had about Yordan Alvarez earlier -- he has an average-size contract with elite hitting metrics, but his listed position is DH.
* **Strikeout rate** (-0.15) — higher strikeout rates may reduce perceived long-term value or offensive stability.

**👨‍💻 Predictive Modeling: Accuracy and Valuations**

[Step 8] The preparation and exploration are behind us, and now we can use what we’ve uncovered so far to predict player contracts. We exported data from [**Baseball Savant**](https://d.docs.live.net/ccd545423950d79d/Documents/MLB_Player_Valuation/baseballsavant.com), used scraping techniques to get contract details from [**Spotrac**](https://www.spotrac.com/mlb/contracts/), career accolades from [Baseball-Reference](https://d.docs.live.net/ccd545423950d79d/Documents/MLB_Player_Valuation/baseball-reference.com), and WAR from [FanGraphs](https://d.docs.live.net/ccd545423950d79d/Documents/MLB_Player_Valuation/fangraphs.com). We did some tedious manual lookups to get Instagram followers. This resulted in a clean dataset with one row per contract/player, with contract-year and career hitting metrics, as well as year-to-year improvements or regressions and first-year to current-year improvements or regressions. We performed exploratory analysis to get a better feel for the data – average metrics for players signing the biggest deals, contract trends, how some variables behave with contract terms. We used k-means clustering to identify player prototypes to help understand which types of players land what kinds of deals. Lastly, we calculated the correlation of all of our predictor variables to our target variables, to see which features correlate most with contract value and length.

Let’s get into modeling.

**📊 Model Preface**

Because this is experimental, I’m not going to get too sophisticated in my modeling approach – I only have 91 observations anyways. The plan is to fit 9 models in total – 4 models for contract length, 4 models for contract value, and a multivariate model that will predict both contract length and value together. In order to do this, I should split my data into a training set and a test set. A fun and logical approach I can take to this is to train on contracts starting prior to 2025, and test on contracts starting in 2025 or later. Hypothetically, what does math suggest is the worth of the players in the most recent free agent class? This gives me 69 observations to train on and 22 observations to test on. What I’ll do from there is calculate accuracy metrics for each, and the model that performs best on unseen data (test set) is the model I’ll move forward with.

* 1. **Multiple Linear Regression**
     + Predict an outcome by adding together effects of several input variables. It assumes a straight-line relationship between each feature (variable) and the outcome (target).
     + Think of this as a scoring system where each stat (WAR, age) contributes a fixed number of points towards a projected contract.
  2. **Lasso Regression**
     + Lasso starts with the same idea as linear regression, but it actively “removes” features that aren’t useful, simplifying the model.
     + You’re packing for a flight with a strict baggage limit — Lasso helps you decide what not to bring by removing unimportant features.
  3. **XGBoost (Extreme Gradient Boosting)**
     + XGBoost builds a series of mini decision trees, where each one tries to fix the mistakes made by the previous ones. Together, they make very accurate predictions.
     + Think of a coach giving feedback after each game — each practice session addresses specific weaknesses from the last one.
  4. **Multivariate Regression**
     + Instead of predicting one outcome (like contract value), multivariate regression predicts multiple related outcomes at the same time (like value and length).
     + This is like forecasting both runs and RBIs for a player — predicting them together because they tend to move in tandem.

**📊 Model Comparison and Selection**

By using accuracy metrics like R² (R-squared) and RMSE (Root Mean Squared Error), I can compare each model’s performance with the others.

* *R²* tells you how well your model explains the variation in the outcome. It ranges from 0 to 1, where **1 means perfect prediction** and **0 means the model is no better than just guessing the average**.
* *RMSE* shows how far off your model’s predictions are, on average. It’s in the **same units as your target variable**, so it’s easy to interpret — lower is better.

Let’s start with the model results on our training set of data:



Figure 18. Model results on training data.

 **XGBoost models achieved the lowest RMSE and highest R² on the training set** for both contract length and value, indicating excellent fit on training data — but this is likely a sign of **overfitting**. Overfitting is when a model learns not just the underlying patterns in the training data, but also the *noise* — random fluctuations or anomalies that don’t generalize to new, unseen data. Despite the rounds of tuning, I could not get the accuracy metrics to drop much further. This is a common issue when using powerful tree-based models on relatively small or structured datasets, and it could also just be that the predictors and contract outcomes are mostly linear in nature.

 **Multivariate linear regression models performed best in predicting both contract length and value. Outside of XGBoost, which I suspect will not generalize well to unseen data (when I want to use this model to actually predict new contracts), the multivariate regression models had the lowest prediction error on average (1.6 years/$53.9M) while also being the best at explaining the variation in outcomes (66.5%/80.7%).**

 **Lasso-Regression had higher RMSE and lower R²** than correlated or multivariate models on training data, suggesting they may have underfit by being too restrictive or dropping important features. However, one benefit to Lasso-regression is that you can leverage it to select the most important features. When fitting a **standard linear model with lasso-selected features**, it **performed better** **than** my **linear models with correlated features**.

Overall, the linear models showed **reasonable training performance without excessive overfitting**, making them a more reliable choice for real-world prediction goals. To confirm this, I need to make predictions on my test set of data.

**📊 Contract Prediction Accuracy on Newest Free Agency Class**

Now that I’ve fit models on my training data, I need to use those models to make predictions on my test data. This will test how well my models perform on unseen data. When fitting the models, the target variables are present in the data so that the model can attempt to reverse-engineer the predictors to calculate the contract length or value. When I use the fitted models on my test set, I’m essentially just using the formulas they calculated on my training data to predict the contract terms. Again, this is necessary because a model performing well on training data doesn’t mean it will generalize better on unseen data (when it can’t use the actual contract terms to help). A quick note: I set the minimum contract length to 1 year and the minimum contract value to $760K.

Let’s see if any models performed better on new data:



Figure 19. Model results on test data.

 **As suspected, performance for XGBoost dropped substantially on the test set, confirming poor generalization to unseen data.**

 **Multivariate linear regression models performed especially well** for contract value again, with predictions being $93M off on average. The model explains 72% of variation in contract value outcomes.

 A **simple linear regression model using the correlated features** predicting contract length performed best on test data. It had an average prediction error of just over 2 years, and the model explained 61% of the variance in contract length outcomes.

When I move on to predicting new contracts for upcoming free agents, I’ll use my **simple linear model with correlated features to predict contract length**, and I’ll use the **multivariate linear model to predict contract value**. First, let’s take a look at some of the model projections for 2025 signees:

A graph with numbers and dots

AI-generated content may be incorrect.

Figure 20. Projecting contracts for most recent free agent class, including most notably Juan Soto.

I used a 25% threshold to classify player valuations. If the projected contract value was more than 25% over the actual contract value, the player was labeled as “undervalued”. If the projected contract value was more than 25% under the actual contract value, the player was labeled as “overvalued”, and if it was within 25%, the contract was labeled as “fair value”. Six players were labeled ‘fair value’ deals, six were ‘overvalued’, and 10 were ‘undervalued’. As a reminder, this class of signees included players like Jose Altuve, Alex Bregman, Juan Soto, Joc Pederson, Pete Alonso, and Vladimir Guerrero Jr.



Figure 21. Count per classification on 2025 signees.

Most accurate predictions (Fair Value):



Figure 22. Accurate projections for 2025 signees, highlighting potential fair value deals.

* The model projected the contract terms for **Jackson Merrill (Cluster 4: High-Floor Contributor)** to the exact length and exact value. On a cluster note, it is surprising to see Jackson in the ‘High-Floor Contributor’ group rather than ‘Young with Upside’. This speaks to his early career success, having played only one season in the big leagues, he made an immediate all-star game appearance, and won a silver slugger award. He had a 5.3 WAR in his first season and has a decent social media following (121K) as well. The Padres are clearly projecting his success to continue, signing him to a 9yr/$135M contract.

Least accurate predictions (Overvalued or Undervalued):



Figure 23. Least accurate projections, possibly highlighting over or undervalued players.

* Let’s call it like it is, **Juan Soto (Cluster 2: Franchise Superstars)** was more than likely blatantly **overpaid** by the New York Mets – but who can blame them if they’ve got the money? It was likely a bidding war for Juan Soto in the 2025 offseason, and rightfully so in my opinion (I’m a Nationals fan, shoutout to 2019). He’s only 25, he is already a 4x All-Star and 5x Silver Slugger, he’s won a batting title and a World Series (again, go Nats). He has elite hitting metrics, and he’s also one of few MLB players with a following of over 1.5M people on Instagram, suggesting his reach off the field (jersey sales, expected increased attendance was surely considered as well). Anyways, my **model** **projected 13 years and $528M**, about $10M less per year than what he is actually making – but two years shorter in length, and about $237M less in overall value.
* 33-year-old **Christian Walker (Cluster 5: Power-first Sluggers)** got 3 years and $60M from the Astros, which the model shows may have been easy to accept from Walker’s standpoint. Obviously nearing the end of his career, he’s never made an All-Star game, but hits HRs at a 5% clip (1 out of every 20 AB) and his performance hasn’t slowed down, boasting a 3.01 WAR in his contract season, almost a point and a half higher than his career average.
* On the other hand, the Astros may have gotten **Jose Altuve (Cluster 1: All-Star Veterans)** on a bargain deal. This could make sense though, maybe he took a hometown discount so the Astros could sign players like Christian Walker. He’s spent his whole career in Houston and is the face of that franchise until he passes the torch to someone else (maybe Jeremy Pena?). The 34-year-old signed a 5yr/$125M contract to end his career in Houston, however my **model** **projected** the hall-of-fame bound utility player should have gotten something like **7 years and $260M** (could he be productive into his 40s?). I think at the end of the day, the deal he signed makes sense for both parties.
* What about SS **Geraldo Perdomo (Cluster 3: Young with Upside)**? Perdomo signed a 4yr/$45M deal with the Diamondbacks but he’s only 24 and already has 1 All-Star game appearance after three seasons in the big leagues. His career metrics suggest a contact-first speedy type of hitter, which may not pay as much in today’s MLB. Nonetheless, my model projected double the contract length for him (8 year deal) in addition to a slightly higher AAV.

**📊 Player Valuations: Overvalued, Undervalued, or Fair Value**

Let’s take a look at our full dataset now, and see if we can spot any outliers or players to highlight:



Figure 24. Count per classification all players.

A graph with different colored dots

AI-generated content may be incorrect.

Figure 25. All player contract projections.



Figure 26. Highest valued contracts per valuation classification.

* The **Ronald Acuna Jr.** deal might be the most team-friendly deal in all of baseball. He has been hurt, but even with the missed time, he’s been well worth the 8yr/$100M contract he signed in 2018. Even at the time of signing, my model projected him to be worth much more than he was signed for (**projected 12yrs/$246M**). If you consider his performance since the contract, the projection probably goes up even higher. Good on the **Braves** for **investing in their young talent early**, as we’ve seen them do with other players as well like Austin Riley and Ozzie Albies.
* The model **accurately projected** big-time contracts like **Mike Trout** and **Aaron Judge** (Franchise Superstars).

**👨‍💻 More Predictive Modeling: What Would it Cost to Sign These Guys Right Now?**

[Step 10] I chose 19 players to project new contracts for, mostly those that are coming up on free agency, or some of my favorite exciting young players. I essentially had to repeat the same data preparation steps I took for my original data and then predict contract length and value with my selected models: **Simple linear regression with correlated features to project contract length**, and a **multivariate regression model to project contract value**.

1. Exported data from [**Baseball Savant**](https://d.docs.live.net/ccd545423950d79d/Documents/MLB_Player_Valuation/baseballsavant.com) and included all stats up to the time of writing this (May 26th, 2025).
2. Aggregated career statistics.
3. Scraped career accolades from [Baseball-Reference](https://d.docs.live.net/ccd545423950d79d/Documents/MLB_Player_Valuation/baseball-reference.com) and WAR from [FanGraphs](https://d.docs.live.net/ccd545423950d79d/Documents/MLB_Player_Valuation/fangraphs.com).
4. Looked up Instagram followers.
5. Normalized data.
6. Predict contracts with fitted models

Here were the results!



A graph with different colored text

AI-generated content may be incorrect.

A chart with different colored dots

AI-generated content may be incorrect.

A graph with different colored dots

AI-generated content may be incorrect.

Figure 27. New player contract projections.

* **Contract Length Predictions** range from 4 to 8 years, indicating a mix of short-term and potential franchise cornerstone deals. **Contract Value Predictions** span widely, with some standout deals above $100 million and others below $30 million.
* **Projected 8-year deals 🡪 Elly De La Cruz, Gunnar Henderson, James Wood, Riley Greene.** Greene is 24, the others are 23, and all are already everyday contributors in the major leagues. **Elly De La Cruz** is an electric switch-hitter that can steal bases on command. With one All-Star appearance already in just three seasons, and a 3.2 career WAR, the Reds will certainly look to make him the face of their franchise. **James Wood** has taken the league by storm in not even a full season’s worth of games played. One of the return pieces for Juan Soto, Wood is forcing the Nats’ hand early. He has elite hitting metrics that most people are confident is not a fluke. He owns an average exit velocity over 93 mph (only Yordan Alvarez and Aaron Judge are higher), and hits HRs at a 5% clip. He’s got 13 home runs in 2025 as of this analysis and is more than likely bound for the All-Star game in his first full season. The model projects the **Nats could sign him right now for 8yrs/$145M**, but that price goes up every day (and he’s a Scott Boras client).
* **Another young National, CJ Abrams**: Forecasted at **7 years, $88.2M**, indicating confidence in his upside despite average performance metrics.
* **Kyle Schwarber** is heading into free agency again and has set himself up for a pretty solid deal to cap off his amazing career that at one point seemed dead as his days in Chicago came to an end. The **model projects the 32-year-old to be worth a 4yr/$145M deal** as of now, leading all of the projected contracts in AAV at over $40M/yr. Aaron Judge and Alex Bregman make a similar amount per year. The model does have a case for projecting this steep price for an older player in Schwarber – he’s not showing many signs of slowing down, with 18 home runs already in 2025, and with the DH option now, it gives teams an option to keep him fresher longer. He’s a 2x All-Star, 1x Silver Slugger, and has a World Series ring, most likely in the ‘**Power-First Slugger**’ cluster.
* **Comparable Contracts:**
  + *Jarren Duran: Projected 5yrs/$50M*
    - JP Crawford, Alejandro Kirk, Brent Rooker
  + *James Wood: Projected 8yrs/$145M*
    - Julio Rodriguez (AAV), Brandon Nimmo, Jose Ramirez
  + *CJ Abrams: Projected 7yrs/$88M*
    - Jake Cronenworth, Byron Buxton, Corbin Carroll, Bryan Reynolds
  + *Kyle Tucker: Projected 7yrs/$177M*
    - Dansby Swanson, Kris Bryant, Willy Adames, Matt Chapman, Marcus Semien
  + *Elly De La Cruz: Projected 8yrs/$208M*
    - Bobby Witt Jr. (AAV), Xander Bogearts (AAV), Trea Turner, Austin Riley

**📝 Closing Remarks**