Predicting 2025-2026 MLB Free Agent Contracts Using R, Clustering, and Regression

Executive Summary:

This project aimed to predict 2025–2026 Major League Baseball (MLB) free agent contracts by analyzing performance and contract data from 2022–2025. Utilizing RStudio, an analytical pipeline was developed that combined k-means clustering, principal component analysis (PCA), and linear regression modeling. The objective was to estimate key contract metrics such as; average annual value (AAV), contract length, and total salary, all based on quantifiable player performance indicators.

Players were segmented using k-means clustering based on standardized performance metrics. PCA was used to reduce dimensionality, and to visualize clusters. Cluster groups were profiled and named according to player types, such as “Very Good Hitters” and “Durable, Effective Starters”.

Linear regression models were then trained using key metrics and cluster labels to predict AAV and years. The models explained 64% of variation in AAV, and 52% in contract length for hitters, with similar performance for pitchers. Variables that had the most significant impact on these models were; WAR, Age, SLG, and vFA..pi. (average fastball velocity). However, the models consistently underpredicted values for elite players, likely due to the limitations of linear regression and the absence of market dynamics such as bidding wars, agent influence, and marketability.

Finally, the trained models were applied to the 2026 free agent class to forecast predicted contracts. Kyle Tucker, the highest coveted upcoming free agent was projected to receive a 7 year $193 million deal. This is a direct example of the model’s linearity, and underrepresents his actual market value. While these forecasts were insightful, it also exposed the limitations of linear regression when it comes to capturing nonlinear salary escalations for elite players.

This report demonstrates how segmentation and modeling can add structure and transparency to MLB contract valuations. While the predictions were largely consistent with real-world outcomes, future improvement such as; integrating non linear models, and market context to refine estimates will be made in the future.

Data Sources and Cleaning:

All performance data were obtained from fangraphs.com and spotrac.com. The initial phase involved downloading hitter statistics from 2022 through April 17, 2025, into an Excel file

and converting it to a CSV format for integration into RStudio. Similarly, a manually compiled spreadsheet of all free agent signings during this period was processed in the same manner.

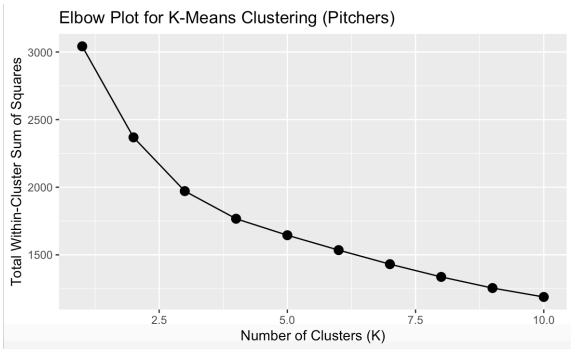
Upon loading both datasets into RStudio, a thorough cleaning process was undertaken. The cleaning process included things such as; converting every name to lowercase, removing all accents, removing all unnecessary columns, removing any N/A values, and finally filtering out any pitchers. The irrelevant columns that were removed primarily consisted of naming variables to identify each player. No columns containing player performance were removed. The datasets were left-joined to create a unified data frame for hitters named "hitters_fa_join." This way, every hitter that signed a free agent deal from 2022-2025, along with their stats from this timeframe was in one data frame.

This cleaning process was repeated for pitchers using identical methods. The resulting dataframe, "pitchers_fa_join" , is the primary data frame when referring to pitchers.

Hitter Segmentation and Methodology:

The following variables were used to create cluster groups, and to predict outcomes; WAR, OBP, SLG, HR, BB., K., SB, AGE, AAV. Initially, the model was run with every offense variable in the data frame. The outcome consisted of conflicting results, due to multicollinearity. Since there were multiple variables that overlapped with each other, it resulted in an unstable and unreliable model. The model that produced the most accurate and logical results was the one consisting of fewer variables that didn't overlap much with one another. Afterwards, the selected variables of the model were standardized to eliminate scale bias.

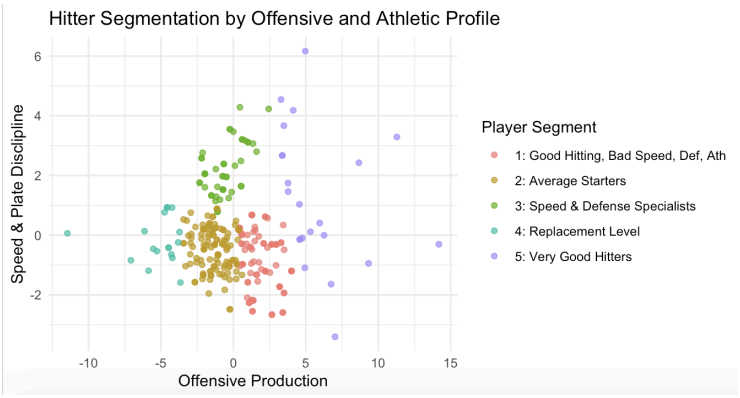
The elbow plot was used to determine an appropriate number of clusters for segmenting hitters. The plot shows a steep drop in total within-cluster sum of squares from 1 to 3 clusters, followed by a more gradual decline. Five clusters were selected to allow for greater differentiation among hitter types. This choice maintained a balance between reducing intra-cluster variance and capturing meaningful variation in offensive profiles.



To enhance interpretability and reduce dimensionality prior to visualizing the cluster groups, a principal component analysis (PCA) was conducted on the standardized variables used in clustering. This technique condensed the input variables into two uncorrelated principal

components, simplifying the structure while preserving the majority of the data's variance. Both of these new components explain the majority of variance across the selected variables, and were both added to the existing data frame. The first principal component (x-axis in the plot) captures overall offensive production. The statistics that influenced this axis the most were; WAR, HR, SLG, OBP, and AAV. All of these variables, excluding AAV, are offensive metrics used to evaluate a hitter's overall performance and skill level. This is why the x-axis was named "Offensive Production". Since PCA1 was inverted, higher values in the associated offensive metrics shift players further to the right along the x-axis. The second principal component (y-axis) captures attributes related to speed, athleticism, and plate discipline. This axis is primarily influenced by stolen bases (SB), walk rate (BB), and strikeout rate (K). Stolen bases reflect a player's speed and athleticism, while BB and K rates indicate plate discipline. Players with stronger performance in these areas are positioned higher on the y-axis. The corresponding plot illustrates the five cluster groups and the contribution of each variable to both principal components.

	PC1	PC2
WAR	-0.44459035	0.16657639
OBP	-0.38957785	-0.21551457
SLG	-0.41506055	-0.12269096
HR	-0.43167782	-0.05975441
BB.	-0.23880711	-0.59518643
K.	0.12530460	-0.39235821
SB	-0.22948236	0.54530722
Age	0.05449848	-0.30475495
AAV	-0.40445668	0.09356671



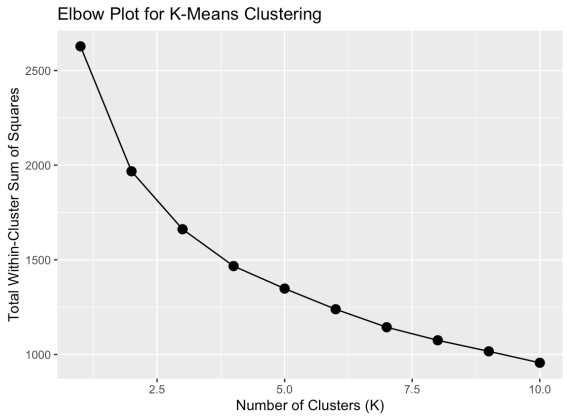
After generating the plot and analyzing the principal component axes, more accurate and descriptive names were assigned to each cluster group. Each cluster was examined based on its statistical profile, allowing for labeling that reflected distinct player archetypes and performance traits.

Pitcher Segmentation and Methodology:

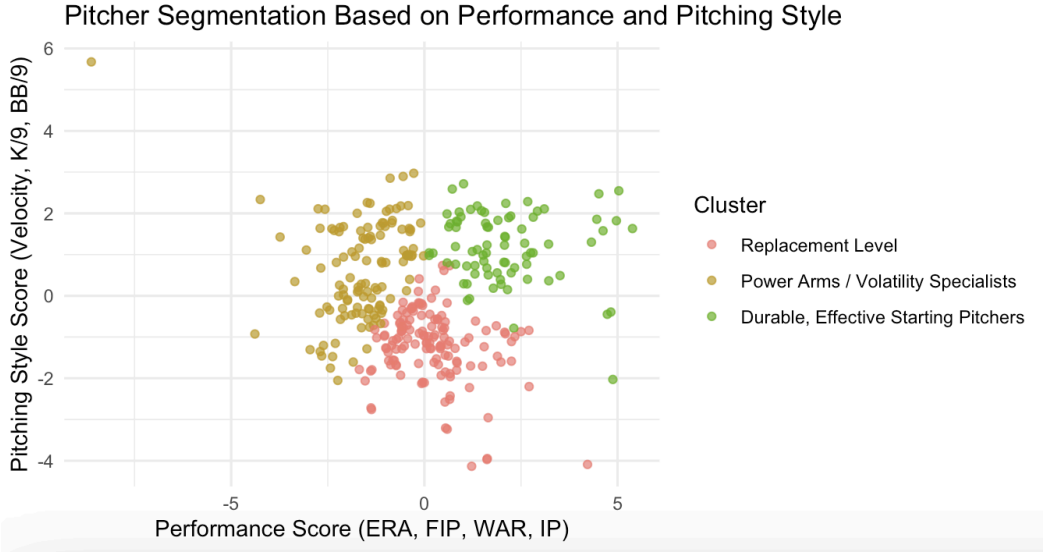
The methodology applied to the pitchers_fa_join dataset followed the same structure as that used for hitters, with adjustments to reflect the unique aspects of pitching performance. Variables used for clustering and outcome prediction included; WAR, ERA, xFIP, innings pitched (IP), strikeouts per nine innings (K/9), walks per nine innings (BB/9), average fastball velocity (vFA..pi.), age, and AAV. Earlier iterations that included a broader set of overlapping metrics produced inconsistent results, likely due to multicollinearity. The final variable set was refined to

include only those that contributed clear, non-redundant information, resulting in more stable and interpretable clusters.

An elbow plot was generated using up to 10 potential cluster groups to evaluate the optimal number of segments for the pitchers_fa_join dataset. Based on the inflection point observed in the plot, three clusters were selected as the most appropriate balance between variance explanation and model simplicity. A set.seed function was applied to ensure the clustering results were reproducible, and three distinct pitcher groups were generated based on the selected input variables.



Principal component analysis (PCA) was conducted on the same standardized variables used for clustering in the pitchers_join_df, following the same approach as with hitters. The first principal component (x-axis in the plot) captured overall pitching performance, with the most influential variables being; WAR, AAV, xFIP, ERA, and IP. Higher-performing pitchers were positioned further to the right along this axis. The second principal component (y-axis) reflected pitching style, influenced primarily by K/9, BB/9, average fastball velocity (vFA..pi.), and IP. This axis was designed to differentiate between power/volatile arms and control-oriented, high-volume pitchers. The y-axis was inverted so that pitchers with high velocity, strikeout rates, and walk rates, traits associated with volatility appeared higher on the plot. Conversely, pitchers lower on the axis tended to feature lower velocity and rate stats but accumulated more innings, reflecting durability and control. The accompanying plot illustrates the three cluster groups and the variable contributions to each principal component.



	PC1	PC2
WAR	-0.47854776	0.27425186
ERA	0.36486076	0.19312111
xFIP	0.41927743	0.26960402
IP	-0.36017803	0.40448432
K.9	-0.28387545	-0.52091454
BB.9	0.18344331	-0.38673167
vFA..pi.	-0.18153826	-0.45102229
Age	0.01742163	0.05523314
AAV	-0.42994245	0.15430852

After generating the PCA plot and examining the distribution along both axes, each cluster group was analyzed in detail to identify defining performance characteristics. This analysis informed the assignment of more accurate and descriptive names to each group, reflecting distinct pitcher archetypes based on statistical profiles and pitching styles.

Hitter Regression Modeling:

Following the creation of cluster groups for the hitters_fa_join dataset, regression models were developed to predict contract outcomes based on player performance from 2022 to 2025 and corresponding free agent signings. To minimize multicollinearity and improve model stability, a reduced set of non-redundant variables was selected; WAR, SLG, HR, Age, BsR, strikeout rate (K.), and Cluster_Name. Two separate linear regression models were constructed, one for predicting average annual value (AAV) and another for contract length (Years).

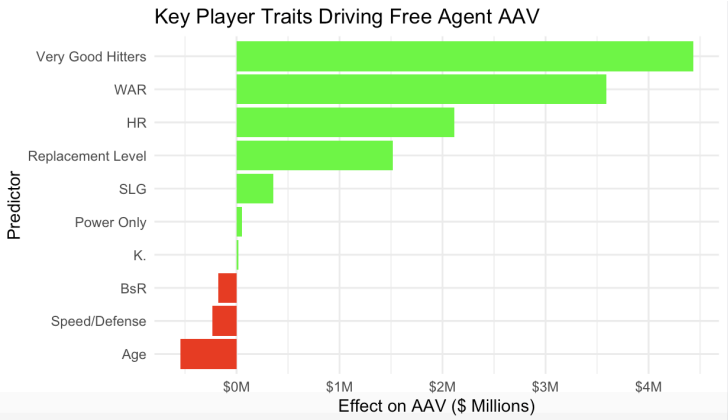
Once both models were finalized and executed, the resulting predictions were added to the hitters_fa_join dataframe. The predicted AAV values were formatted in millions with two decimal places, while predicted years were rounded to whole numbers for clarity. A third column, predicted_total_salary, was then created by multiplying the predicted AAV by predicted contract length; this value was also formatted in millions to maintain consistency across output variables.

1. **Predicted_AAV Model:**

The summary output of the predicted AAV model produced an adjusted R-squared value of 0.639, indicating that approximately 63.9% of the variation in average annual value across players was explained by the independent variables; WAR, SLG, HR, Age, BsR, strikeout rate (K), and cluster group. This level of explanatory power suggests the model effectively captures a substantial portion of the underlying factors influencing AAV. While no model in the context of sports economics is expected to be perfect due to the complexity of contract negotiations, the result is both statistically meaningful and practically useful. This is further supported by an F-statistic of 52.66, confirming the overall significance of the regression model. The following analysis explores the influence of each variable on the model's output.

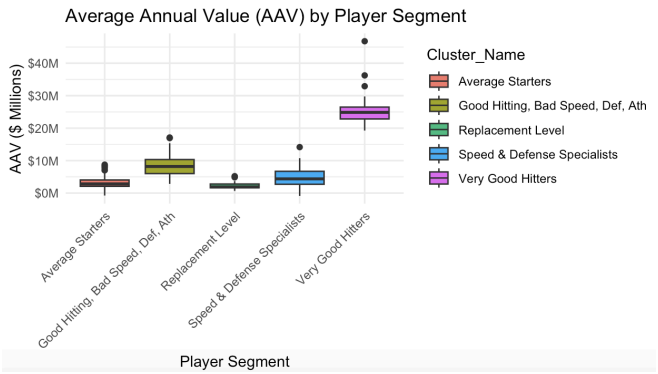
```
Call:
lm(formula = AAV ~ WAR + SLG + HR + Age + BsR + K. + Cluster_Name,
    data = hitters_standardized)

Residuals:
      Min        1Q    Median        3Q       Max
-10258616  -2708639   -973105   1998016  24176356

Coefficients:
                                             Estimate Std. Error t value
(Intercept)                                   6209155     568410  10.924
WAR                                           3589584     631779   5.682
SLG                                            356472     616216   0.578
HR                                            2114476     684134   3.091
Age                                           -547419     305124  -1.794
BsR                                           -177903     422104  -0.421
K.                                             18201     346832   0.052
Cluster_NameGood Hitting, Bad Speed, Def, Ath    51950     979320   0.053
Cluster_NameReplacement Level                 1514428    1710996   0.885
Cluster_NameSpeed & Defense Specialists       -234562    1101730  -0.213
Cluster_NameVery Good Hitters                 4434553    2285520   1.940
                                             Pr(>|t|)
(Intercept)                                   < 2e-16 ***
WAR                                          3.32e-08 ***
SLG                                            0.5634
HR                                             0.0022 **
Age                                            0.0739 .
BsR                                            0.6737
K.                                             0.9582
Cluster_NameGood Hitting, Bad Speed, Def, Ath   0.9577
Cluster_NameReplacement Level                  0.3768
Cluster_NameSpeed & Defense Specialists        0.8316
Cluster_NameVery Good Hitters                  0.0533 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4940000 on 282 degrees of freedom
Multiple R-squared:  0.6512,    Adjusted R-squared:  0.6389
F-statistic: 52.66 on 10 and 282 DF,  p-value: < 2.2e-16
```

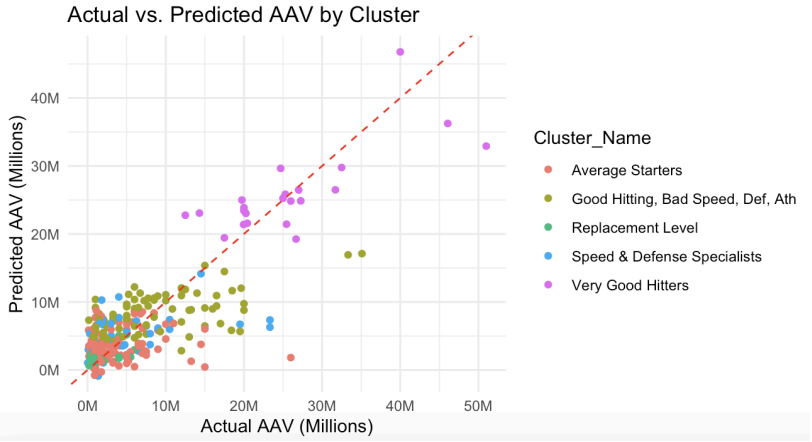
This plot demonstrates what independent variables influence free agent AAV. The predictor that had the largest impact on AAV was the “Very Good Hitters” cluster group with a p-value of 0.0533.. Players that belong to this segment saw their AAV rise by $4.43 million, in reference to the baseline “Average Starters” group). WAR emerged as the most statistically significant predictor in the model, with a p-value of 3.32e-08, indicating a strong and reliable relationship between a player's WAR and their predicted average annual value.



Each 1 unit increase in WAR, results in a raise in AAV by $3.59 million. The last statistically significant contributor was HR, with a p-value of 0.0022. Each home run increases AAV by $2.1 million, as it’s conditional on other factors. The last variable that had any significance was Age. With a p-value of 0.0739, it’s demonstrated that with each year a player ages, they tend to lose $574,419 in AAV. This makes logical sense, as the older a player is, the more likely they will earn annually due to their age. Another variable worth highlighting is the “Replacement Level” cluster group. This is the worst cluster group to belong to (performance wise); however, these players are predicted to earn about $1.5 million more than the baseline group. The majority of players in this cluster group are aging veterans who’s best playing days are behind them. While they may not offer much production on the field at this point, they are still valuable assets to the teams signing them. Serving as leaders and experienced veterans are two qualities we aren’t able to quantify in terms of salary. While this group tends to make more money, it’s more than likely due to chance, along with previous factors stated above.

This box plot demonstrates the average predicted AAV by cluster group. It should come as no surprise to see the “Very Good Hitters” group dominate this visualization. The average predicted AAV for this cluster was $26 million, while the next closest cluster group was “Good Hitting, Bad Speed, Def, Ath” coming in at a predicted AAV of $8 million.





The final plot in this section illustrates the overall accuracy of the model. The x-axis represents actual AAV values, while the y-axis shows the model's predicted AAV. Each point corresponds to a player in the hitters_fa_join dataset, providing a visual representation of prediction accuracy relative to real-world contract figures. The model demonstrates greater precision in forecasting lower AAV values, while predictions become less reliable as AAV increases, indicating a tendency to underpredict contracts for higher-value players.

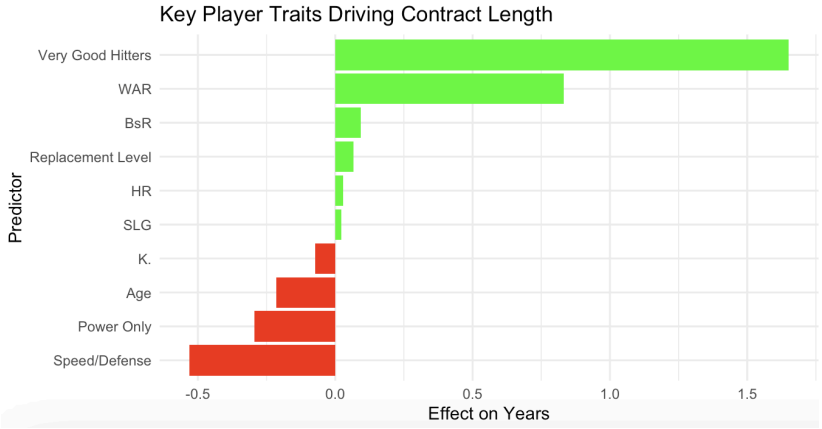
2. Predicted_Years Model:

After running a summary of the predicted_years model, the adjusted r-squared score was 0.5156. This means that about 51.6% of the variation in Years across players is explained by the independent variables in the model (same variables as before). Essentially, the model captures nearly 52% of underlying patterns that determine the years of a contract. While it's not as strong as the AAV model, it still demonstrates a solid fit. The F-statistic of this model is 32.08. Below is a summary of the model, with an analysis of what variables affected it.

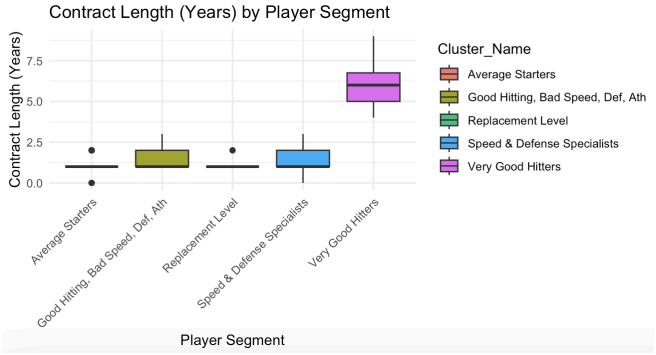
```
Call:
lm(formula = Years ~ WAR + SLG + HR + Age + BsR + K. + Cluster_Name,
    data = hitters_standardized)

Residuals:
    Min      1Q  Median      3Q     Max
-4.1616 -0.5205 -0.1335  0.1689  7.8118

Coefficients:
                                           Estimate Std. Error t value
(Intercept)                                 1.71765    0.14610  11.756
WAR                                         0.83315    0.16239   5.130
SLG                                         0.02139    0.15839   0.135
HR                                          0.02927    0.17585   0.166
Age                                        -0.21344    0.07843  -2.721
BsR                                         0.09331    0.10850   0.860
K.                                         -0.07314    0.08915  -0.820
Cluster_NameGood Hitting, Bad Speed, Def, Ath -0.29425    0.25172  -1.169
Cluster_NameReplacement Level               0.06729    0.43979   0.153
Cluster_NameSpeed & Defense Specialists    -0.52999    0.28319  -1.871
Cluster_NameVery Good Hitters               1.65093    0.58747   2.810
                                           Pr(>|t|)
(Intercept)                                 < 2e-16 ***
WAR                                        5.39e-07 ***
SLG                                           0.8927
HR                                            0.8679
Age                                           0.0069 **
BsR                                           0.3905
K.                                            0.4127
Cluster_NameGood Hitting, Bad Speed, Def, Ath   0.2434
Cluster_NameReplacement Level                 0.8785
Cluster_NameSpeed & Defense Specialists       0.0623 .
Cluster_NameVery Good Hitters                 0.0053 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.27 on 282 degrees of freedom
Multiple R-squared:  0.5322,    Adjusted R-squared:  0.5156
F-statistic: 32.08 on 10 and 282 DF,  p-value: < 2.2e-16
```

This plot to the bottom right of this, demonstrates what independent variables influence free agent contract length (Years). The predictor that had the largest impact on Years was the "Very Good Hitters" cluster group with a p-value of 0.0053. This should come as no surprise, based on the AAV plot. Players belonging to this group saw an average of about 1.6 additional years on their contract, in reference to the baseline group. Once again, WAR stands out

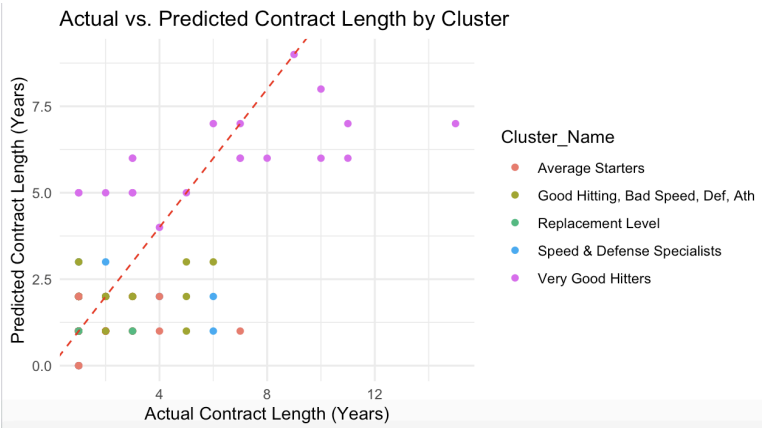
as the most statistically significant predictor, with a p-value of 5.39e-07. Each 1 unit increase in WAR, gives an additional 0.833 years to a player's contract, holding all other variables constant. Age was also a statistically significant variable, with a p-value of 0.0069. As age increases, players can expect about 0.20 years off of their contract each year. The last somewhat statistically significant variable was the "Speed & Defense Specialists" cluster group, with a p-value of 0.0623. While it's just outside of the statistically significant threshold, it still indicates some evidence of an effect on predicted years. Many of the players in this group play multiple positions, which is a skill that wasn't accounted for in the model. This leads to a possible market undervaluation of this cluster group in the free agent market.



This box plot demonstrates the average predicted years by cluster group. It again shouldn't be shocking to see the "Very Good Hitters" group outperform the other groups. The average predicted years for this cluster was 6 years, while the next closest cluster group was "Good Hitting, Bad Speed, Def, Ath" coming in at predicted years of 2.



The final plot in this section displays the overall accuracy of the contract length model. The x-axis represents actual contract years, while the y-axis shows the model's predicted values. Because contract duration is measured in whole numbers, the plot contains fewer distinct points, with many overlapping. Nevertheless, it



provides a clear visual representation of the model's predictive alignment with real-world free agent signings.

Pitcher Regression Modeling:

After generating cluster groups for the pitchers_fa_join dataset, regression models were developed to predict contract terms based on player performance from 2022 to 2025 and corresponding free agent signings. The modeling process followed the same structure as that used for hitters, with one key adjustment: the variable "cluster_name" was excluded. Including it in earlier model iterations produced illogical or inconsistent results, prompting its removal to improve model coherence.

Following the construction and execution of both the predicted AAV and predicted years models, the corresponding columns were added to the pitchers_fa_join dataframe. As with hitters, the predicted_total_salary column was calculated by multiplying predicted AAV by predicted contract length, and all monetary values were converted to millions with two decimal places for consistency. The variables used in these models included WAR, ERA, xFIP, K/9, BB/9, IP, average fastball velocity (vFA..pi.), and Age.

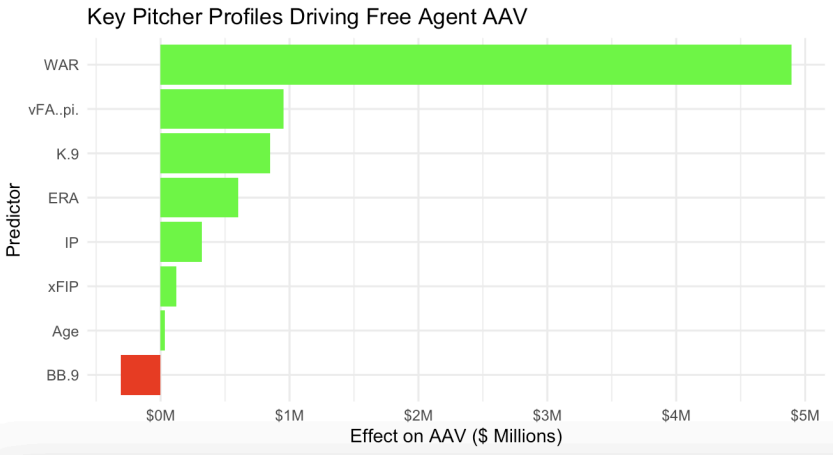
Predicted_AAV Model:

After running a summary of the predicted_aav_p model, the adjusted r-squared score was 0.5345. This means that about 53% of the variation in AAV across players is explained by the independent variables in the model. Essentially, the model captures nearly 53% of underlying patterns that determine AAV. Similar to the hitters_fa_join predicted years model, it demonstrates a solid fit. The f-statistic of this model is 49.51. Below is a summary of the model, and with an analysis of what variables affected it.

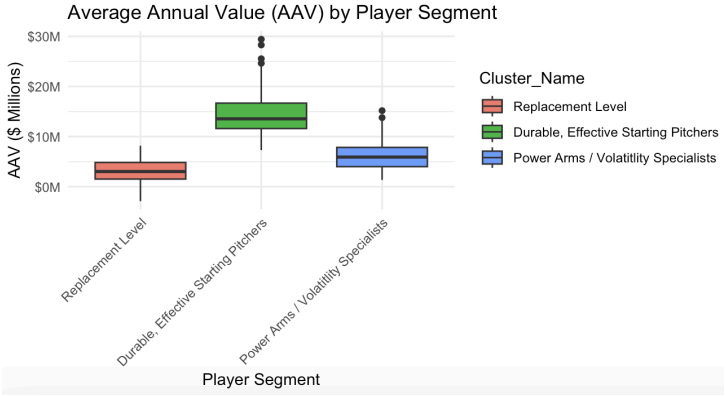
```
Call:
lm(formula = AAV ~ WAR + ERA + xFIP + K.9 + BB.9 + IP + vFA..pi. +
    Age, data = pitchers_standardized)

Residuals:
      Min        1Q    Median        3Q       Max
-11429980  -2953878   -643651   2120816  26688517

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7233911     267680  27.024  < 2e-16 ***
WAR          4893272     585461   8.358  1.8e-15 ***
ERA           602988     341396   1.766  0.07828 .
xFIP          124228     534638   0.232  0.81640
K.9           851422     505753   1.683  0.09323 .
BB.9         -307158     400659  -0.767  0.44385
IP            320735     541504   0.592  0.55405
vFA..pi.      953136     324307   2.939  0.00352 **
Age            33561     280224   0.120  0.90474
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4929000 on 330 degrees of freedom
Multiple R-squared:  0.5455,    Adjusted R-squared:  0.5345
F-statistic: 49.51 on 8 and 330 DF,  p-value: < 2.2e-16
```

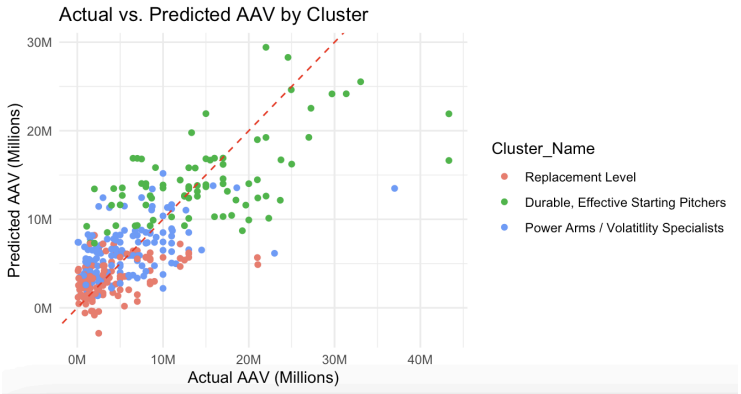
The plot to the bottom right of this, demonstrates what independent variables influence free agent AAV. The predictor that had the largest impact on AAV was WAR, with the highest statistically significant value of 1.8e-15. Each 1 unit increase in WAR, results in a raise in AAV by $4.89

million. The other statistically significant variable was vFA..pi. (average fastball velocity), with a p-value of 0.0035. Each MPH on a pitcher's fastball results in an additional $953,000 in AAV. xFIP, BB/9, IP, and Age were all non-significant statistically. While they weren't statistically significant, K/9 with a p-value of 0.0932, and ERA with a p-value of 0.0783, both resulted in $600,00-800,000 on AAV. Both of these variables may slightly influence AAV; however, not with high statistical confidence. Thanks to this plot, we can see how much teams are valuing WAR and vFA..pi.. Over the past decade, Major League Baseball has experienced a significant increase in average fastball velocity among pitchers. This shift, often referred to as the "velocity revolution," has made fastball velocity a critical component of pitcher valuation. It is unsurprising that pitchers who throw harder tend to command higher salaries. Given current league trends and organizational preferences, this emphasis on velocity is expected to persist well into the future.



This box plot illustrates the average predicted AAV across the three pitcher cluster groups. The "Durable, Effective Starting Pitchers" cluster appears at the top, with an average predicted AAV of $14 million. In comparison, the "Power Arms / Volatility Specialists" cluster follows at a significantly lower average of $6 million. This clear gap highlights how consistency, workload, and overall effectiveness heavily influence projected contract value for pitchers.



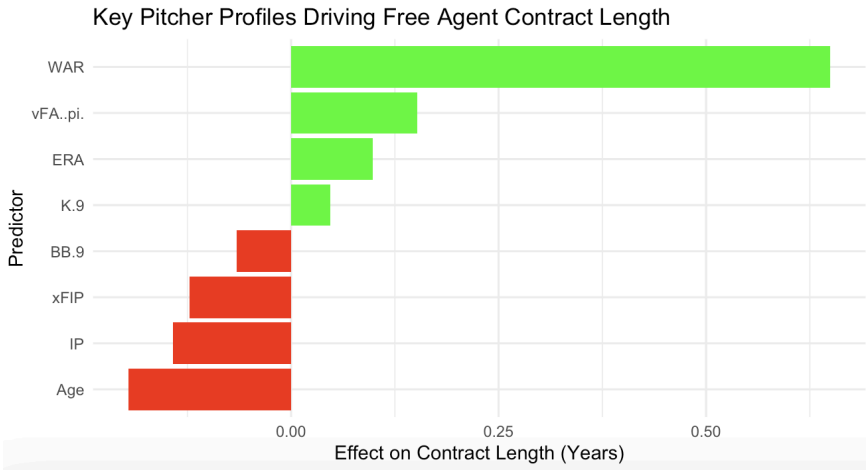


Finally, the last plot for this section is the overall model accuracy. The x-axis shows actual AAV, while the y-axis is what the model predicted. Every dot in the plot represents each player in the pitchers_fa_join df. This visualization provides a clear sense of the model's predictive accuracy relative to actual free agent signings. Similar to the hitters_fa_join dataset, the model demonstrates greater precision when estimating lower AAV values. However, once pitcher AAVs exceed approximately $10 million, there is a noticeable increase in variability between predicted and actual values, reflecting the model's reduced accuracy at the higher end of the market.

1. Predicted_Years Model:

After running a summary of the predicted_years model, the adjusted r-squared score was 0.3824. This means about 38% of the variation in Years across players is explained by the independent variables in the model. Essentially, the model captures nearly 38% of underlying patterns that determine the years of a contract. This model obviously needs improvement, and could be subject to future engineering. The F-statistic of this model is 27.16. Below is a summary of the model, with an analysis of what variables affected it.

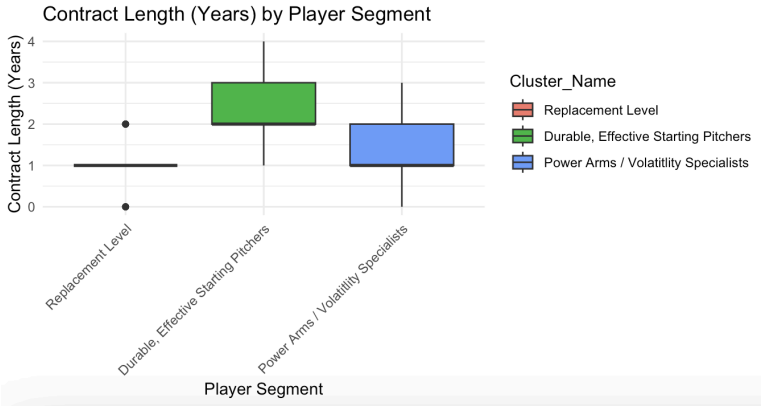
```
Call:
lm(formula = Years ~ WAR + ERA + xFIP + K.9 + BB.9 + IP + vFA..pi. +
    Age, data = pitchers_standardized)

Residuals:
    Min      1Q  Median      3Q     Max
-2.1679 -0.4895 -0.1235  0.3290  4.3603

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.55457    0.04560  34.092  < 2e-16 ***
WAR          0.64988    0.09973   6.516 2.70e-10 ***
ERA          0.09842    0.05816   1.692  0.09153 .
xFIP        -0.12279    0.09107  -1.348  0.17853
K.9          0.04736    0.08615   0.550  0.58291
BB.9        -0.06537    0.06825  -0.958  0.33889
IP          -0.14245    0.09224  -1.544  0.12347
vFA..pi.     0.15211    0.05525   2.753  0.00623 **
Age         -0.19625    0.04774  -4.111 4.97e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8396 on 330 degrees of freedom
Multiple R-squared:  0.397,     Adjusted R-squared:  0.3824
F-statistic: 27.16 on 8 and 330 DF,  p-value: < 2.2e-16
```



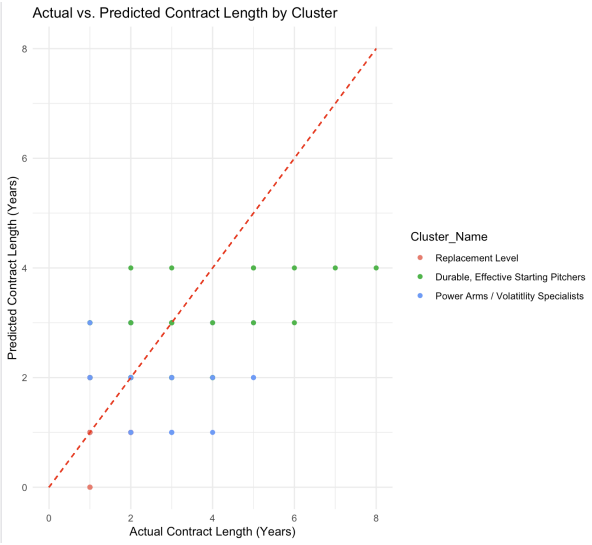
This plot demonstrates what independent variables influence free agent contract length (Years). The predictor with the largest impact on Years was WAR with a p-value of 2.70e-10. Each additional increase in WAR adds 0.65 years to a pitcher's contract length, holding other factors constant. vFA..pi. was once again statistically significant, with a p-value of 0.0062. Each MPH

increase on a pitcher's fastball velocity results in 0.15 more years on average. The last variable that was statistically significant was age, with a p-value of 4.97e-05. This is a negative relationship, as each additional year of age reduces contract length by 0.20 years. xFIP, K/9, BB/9, and IP were all not statistically significant. ERA with a p-value of 0.0915 was marginally significant, but not as significant as the other variables. WAR, vFA..pi., and age are all key predictors of how long a pitcher's contract will be. It's clear that teams don't mind paying older pitchers a bit more annually (AAV), but they will not give them a long-term deal, due to their age and injury risk.

This box plot demonstrates the average predicted years by cluster group. "Durable, Effective Starting Pitchers" lead this visualization; however, not by much. Both this cluster, and the "Power Arms / Volatility Specialists" clusters have an average predicted years of 2.



Finally, the last plot for this section is the overall model accuracy. The x-axis shows actual contract length, while the y-axis is what the model predicted. Since contract length is restricted to whole numbers, the plot contains fewer distinct points and a higher degree of overlap compared to the AAV visualization. Despite this, the plot still offers a clear visual representation of the model's accuracy in predicting contract length based on actual free agent signings.



2026 Free Agent Predictions Hitters:

To obtain the upcoming class of free agent position players, a list including player names and ages was copied from Spotrac and imported into RStudio as a CSV file titled "fa_2026_hitters." The data cleaning process involved renaming and removing unnecessary columns, converting all names to lowercase, stripping accents, and eliminating any rows with missing values. The cleaned dataset was then joined with the existing "hitters_clean_2022_2025" dataframe, resulting in a new dataframe,

“fa_2026_hitters_stats”. This contained each projected 2026 free agent hitter along with their performance statistics from 2022 to 2025.

With the newly cleaned dataframe in place, the next step was to conduct cluster analysis using the same methodology applied earlier. The same set of offensive performance variables was selected to maintain consistency in segmentation. After standardizing these variables, cluster groups were assigned to the “fa_2026_hitters_stats” dataframe. Once clustering was complete, the process transitioned to predictive modeling to estimate contract terms for each player.

Following the established modeling process, the regression models were used to predict each player’s AAV and contract length based on the selected offensive variables. The predicted_years column was rounded to whole numbers, and a predicted_total_salary column was calculated by multiplying predicted_aav by predicted_years. Both monetary columns, predicted_aav and predicted_total_salary were formatted in millions with two decimal places for clarity. After reorganizing the data frame columns for presentation, the final version of the fa_2026_hitters_stats dataset was ready for review.

The fa_2026_hitters_stats data frame is displayed to the right, filtered to show predicted AAV values from highest to lowest. This view highlights the top 33 projected free agent deals based on predicted AAV. While the full dataset includes each player's performance statistics from 2022 to 2025, those columns were excluded from the image to maintain focus on the predicted contract outcomes.

	Name_clean	Age	Predicted_Years	Predicted_Total_Salarly	Predicted_AAV	Cluster_Name
93	kyle tucker	29.2	7	193.05	27.58	Very Good Hitters
61	alex bregman	32.1	6	156.53	26.09	Very Good Hitters
71	pete alonso	31.3	5	127.43	25.49	Very Good Hitters
98	william contreras	28.3	6	147.15	24.52	Very Good Hitters
50	kyle schwarber	33.1	4	92.02	23.01	Very Good Hitters
4	paul goldschmidt	38.6	5	109.35	21.87	Very Good Hitters
26	marcell ozuna	35.4	4	84.27	21.07	Very Good Hitters
33	j.t. realmuto	35.1	3	45.27	15.09	Good Hitting, Bad Speed, Def, Ath
99	bo bichette	28.1	3	41.87	13.96	Good Hitting, Bad Speed, Def, Ath
97	josh naylor	28.8	3	41.75	13.92	Good Hitting, Bad Speed, Def, Ath
90	gleyber torres	29.3	3	41.46	13.82	Good Hitting, Bad Speed, Def, Ath
80	cody bellinger	30.8	3	39.59	13.20	Good Hitting, Bad Speed, Def, Ath
82	ha-seong kim	30.5	3	38.92	12.97	Speed & Defense Specialists
96	luis robert jr.	28.8	2	25.56	12.78	Speed & Defense Specialists
67	cedric mullins	31.6	3	37.22	12.41	Speed & Defense Specialists
22	max muncy	35.7	2	23.80	11.90	Good Hitting, Bad Speed, Def, Ath
91	ozzie albies	29.2	3	35.11	11.70	Good Hitting, Bad Speed, Def, Ath
88	jarren duran	29.6	3	33.62	11.21	Speed & Defense Specialists
81	lane thomas	30.7	2	19.88	9.94	Speed & Defense Specialists
40	joc pederson	34.0	2	19.65	9.82	Good Hitting, Bad Speed, Def, Ath
65	brandon lowe	31.8	2	19.48	9.74	Good Hitting, Bad Speed, Def, Ath
58	lourdes gurriel jr.	32.5	2	18.20	9.10	Good Hitting, Bad Speed, Def, Ath
95	luis arraez	29.0	3	27.31	9.10	Good Hitting, Bad Speed, Def, Ath
2	carlos santana	40.0	1	9.00	9.00	Good Hitting, Bad Speed, Def, Ath
79	tyler o'neill	30.8	2	17.97	8.98	Good Hitting, Bad Speed, Def, Ath
84	thairo estrada	30.2	2	17.90	8.95	Speed & Defense Specialists
18	salvador perez	35.9	1	8.72	8.72	Good Hitting, Bad Speed, Def, Ath
21	mike yastrzemski	35.7	2	17.02	8.51	Good Hitting, Bad Speed, Def, Ath
76	danny jansen	31.0	2	16.51	8.26	Average Starters
47	max kepler	33.2	2	16.29	8.15	Good Hitting, Bad Speed, Def, Ath
94	willi castro	29.0	2	16.24	8.12	Speed & Defense Specialists
42	brandon drury	33.7	1	7.84	7.84	Good Hitting, Bad Speed, Def, Ath
92	luis rengifo	29.2	2	15.39	7.69	Speed & Defense Specialists

Showing 1 to 33 of 99 entries, 23 total columns

It’s important to note that not all player listed are

guaranteed to reach free agency following this season. Many players in this image such as; Bregman, Alonso, Bellinger, Pederson, Durran, etc may not reach free agency due to contract clauses such as; opt-outs, club options, mutual options, or arbitration eligibility. This model and data frame is predicting what all of these players would get on the open market if they do in fact reach free agency.

It comes to no surprise for baseball fans that Kyle Tucker is at the top of this free agent class, with a projected 7 years $193 million deal ($27.58 AAV). Tucker has been one of the best players in all of baseball over this time frame, and is just 29 years old. While the model does a fair job of estimating contract length and AAV across most players, it consistently underpredicts the upper bound of elite free agents. This underestimation can be explained through two key limitations.

The first reason is because the model is a linear model. Linear models tend to flatten predictions around the average. This tendency causes two main effects, the first being underprediction for extreme values. Since the model assumes the relationship remains linear across the full distribution, it will value a jump from a 2 to 3 WAR player, the same as a 6 to 7 jump. This is significant because a player that jumps from 2 to 3 WAR may see a 25% increase in AAV. Meanwhile, a player that jumps from 6 to 7 WAR might see a 50-75% increase in AAV, due to; scarcity, perceived star potential, and estimated future production. Another pitfall of the linear regression model is the overprediction for below-average players. Since the model attempts to minimize total error across all observations and variables, it typically predicts these lower-level players at a higher range than they would typically receive.

Another major reason for the undervaluation of elite players is due to uncaptured market dynamics. The statistical model used only captures quantitative performance and contract metrics. Elements such as; team-specific goals, bidding wars, agent influence, player marketability, positional scarcity, and fan appeal don’t influence the model whatsoever. Bidding wars among teams are a common occurrence during each free agency period. Additionally, players that are marketable and appeal to the fan’s demographic, command more money in free agency than the average player. Since these qualities are intangible, the model cannot take them into consideration. While a franchise may be overpaying on paper for a marketable player, it will likely work out for the front office in the long run, due to the revenue this player creates.

While the current model captures a substantial portion of the free agent market, it consistently underpredicts contract values for elite players due to the limitations of linear regression and the exclusion of key external variables. Future improvements could include the development of non-linear models such as piecewise regression, exponential or power models,

random forest, and XGBoost. These approaches are better equipped to account for the disproportionate market value assigned to high-performing players. Additionally, integrating economic and negotiation-related variables; such as team payroll flexibility, agent reputation, market size, and positional scarcity would enhance the model's predictive power, particularly at the top end of the market. While the existing model offers reliable insights across much of the player pool, these refinements would significantly improve its accuracy for projecting elite free agent deals.

2026 Free Agent Predictions Pitchers:

For the upcoming free agent pitcher predictions, the same process used for the fa_2026_hitters_stats data frame was applied. After loading, cleaning, and preparing the necessary datasets, a new dataframe titled “fa_2026_pitchers_stats” was created. This was generated by joining the newly compiled fa_2026_pitchers data frame with the existing pitchers_clean_2022_2025 data frame. The resulting data frame includes each projected 2026 free agent pitcher, along with their complete performance statistics from 2022 through 2025.

After cleaning the new data frame, the same set of variables used in the earlier analysis was standardized for consistency. Cluster groups were then assigned to the fa_2026_pitchers_stats data frame based on these standardized variables. Once segmentation was complete, predictive modeling was applied following the same structure as with the hitters data frame. The predicted_aav and predicted_years outputs were used to calculate predicted_total_salary. All monetary values were rounded to two decimal places in millions, and contract lengths were rounded to whole years. After final formatting and reordering of columns, the data frame was ready for presentation.

	Name_clean	Age	Predicted_Years	Predicted_Total_Salarly	Predicted_AAV	Cluster_Name
126	dylan cease	30.3	4	107.92	26.98	Durable, Effective Starting Pitchers
100	framber valdez	32.4	4	97.48	24.37	Durable, Effective Starting Pitchers
123	zac gallen	30.8	4	94.80	23.70	Durable, Effective Starting Pitchers
1	justin verlander	43.2	2	40.08	20.04	Durable, Effective Starting Pitchers
26	chris sale	37.1	3	56.28	18.76	Durable, Effective Starting Pitchers
105	zach eflin	32.0	3	53.97	17.99	Durable, Effective Starting Pitchers
25	chris bassitt	37.2	2	34.72	17.36	Durable, Effective Starting Pitchers
119	michael king	30.9	3	51.39	17.13	Durable, Effective Starting Pitchers
20	miles mikolas	37.7	2	33.48	16.74	Durable, Effective Starting Pitchers
132	freddy peralta	29.8	3	49.98	16.66	Durable, Effective Starting Pitchers
34	seth lugo	36.4	2	30.50	15.25	Durable, Effective Starting Pitchers
81	jordan montgomery	33.3	2	30.44	15.22	Durable, Effective Starting Pitchers
3	max scherzer	41.8	2	30.22	15.11	Durable, Effective Starting Pitchers
108	ryan helsley	31.8	3	43.44	14.48	Power Arms / Volatitlity Specialists
115	nestor cortes	31.3	2	28.72	14.36	Durable, Effective Starting Pitchers
120	shane bieber	30.9	3	42.60	14.20	Durable, Effective Starting Pitchers
21	merrill kelly	37.5	2	27.74	13.87	Durable, Effective Starting Pitchers
70	jon gray	34.4	2	25.28	12.64	Durable, Effective Starting Pitchers
14	alex cobb	38.5	2	25.18	12.59	Durable, Effective Starting Pitchers
2	charlie morton	42.4	1	12.54	12.54	Durable, Effective Starting Pitchers
16	clayton kershaw	38.1	2	24.98	12.49	Durable, Effective Starting Pitchers
36	tyler anderson	36.3	2	24.76	12.38	Durable, Effective Starting Pitchers
13	kyle gibson	38.5	2	24.72	12.36	Durable, Effective Starting Pitchers
23	jose quintana	37.2	2	24.14	12.07	Durable, Effective Starting Pitchers
83	brandon woodruff	33.2	2	24.14	12.07	Power Arms / Volatitlity Specialists
125	jack flaherty	30.5	2	23.92	11.96	Durable, Effective Starting Pitchers
65	andrew heaney	34.8	2	22.96	11.48	Durable, Effective Starting Pitchers
111	devin williams	31.6	2	22.60	11.30	Power Arms / Volatitlity Specialists
46	nick martinez	35.7	2	22.28	11.14	Durable, Effective Starting Pitchers
69	matt strahm	34.4	2	22.28	11.14	Power Arms / Volatitlity Specialists
60	marcus stroman	35.0	2	22.00	11.00	Durable, Effective Starting Pitchers
15	aroldis chapman	38.2	2	20.94	10.47	Power Arms / Volatitlity Specialists
4	chris martin	39.9	2	20.22	10.11	Power Arms / Volatitlity Specialists

Showing 1 to 33 of 135 entries, 21 total columns

The fa_2026_pitchers_stats

data frame is displayed above. It highlights the top 33 projected free agent deals for pitchers based on model predictions. As with the hitters' forecasts, not all listed players are guaranteed to enter free agency, as certain contractual clauses may exist.

Analyzing the fa_2026_pitchers_stats data frame, the model appears to have performed well in identifying the top projected earners among upcoming free agent pitchers. Most of the high-ranking predictions align with expected market values based on recent performance. One notable exception is Justin Verlander, whose predicted AAV places him among the top five. While this may not align with subjective expectations, sustained performance at his current level could justify such a valuation. Although the model predicts multi-year deals for aging veterans like Verlander and Max Scherzer, projections that may be optimistic given their age and durability concerns, it generally produces reasonable estimates for contract length across the broader pitcher pool.

Another major effect we see on the predicted contract values is due to "cluster_name". Obviously, "Durable, Effective Starting Pitchers" dominate the majority of the top 33 spots. The majority of pitchers that fall into the "Power Arms / Volatility Specialists" are relief pitchers. It appears the model did a great job of risk-adjusted predictions for age and volatility. Older pitchers along with more injury prone pitchers have lower AAV, despite strong track records.

While this model also accurately reflects the majority of the pitching free agent market, it still slightly underpredicted elite free agent pitchers, similar to the fa_2026_hitters_stats data frame. This section of the project can be improved with the same factors mentioned earlier (non-linear regression models and economic/negotiation variables).

Findings on Past Free Agent Contracts:

Two notable insights emerged during the model generation process, specifically related to past free agent contracts. This data frame, represents every 2022-2025 position player signing, from the most overpaid to underpaid contracts. Here, we will dissect those who are overpaid, according to the

	Player	Year	Actual AAV (M)	Predicted AAV (M)	Diff (M)
1	kris bryant	2022	26.00	1.82	-24.18
2	juan soto	2025	51.00	32.92	-18.08
3	carlos correa	2022	35.10	17.11	-17.99
4	javier báez	2022	23.33	6.29	-17.04
5	carlos correa	2023	33.33	16.93	-16.40
6	trevor story	2022	23.33	7.35	-15.98
7	nelson cruz	2022	15.00	0.44	-14.56
8	josé abreu	2023	19.50	5.67	-13.83
9	starling marte	2022	19.50	6.72	-12.78
10	brandon belt	2022	18.40	5.83	-12.57
11	avisaíl garcía	2022	13.25	1.28	-11.97
12	anthony rizzo	2023	20.00	8.80	-11.20
13	mitch haniger	2023	14.50	3.77	-10.73
14	nick castellanos	2022	20.00	9.75	-10.25
15	michael conforto	2025	17.00	6.82	-10.18

model. The majority of deals on this list shouldn't come as a surprise. These represent bad deals, along with poor performance the franchise may not have foreseen. A good amount of these players have also struggled with injury issues. While that may not entirely be their fault, it still results in bad signings for that franchise.

There's one name on this list that is an exception to poor performance, injury issues, and overall bad deals. Juan Soto signed an unprecedented 15 year, $765 million ($51 million AAV) this past offseason. There's no denying he's a top 5 consensus hitter in the game; however, many baseball fans including myself considered this an overpay. While Soto's entire hitting skillset is generational, that's quite literally all he brings to the plate. Soto offers poor to average defense and speed at best. While these two skills are nowhere near as important as offensive production, they still play a major role in free agent contracts. Due to Soto's poor speed and defense, many envision him moving off the outfield into a DH role in the future. I'd personally be shocked if Soto plays 10 of his 15 years in the outfield. DH's are valued contractually, and analytically much less than position players due to their one-dimensional impact on the game. Additionally, 10+ year deals are always a gamble. Soto will be 41 in year 15 of his contract. If his production falls drastically due to age and general decline, this may turn into a horrible contract. While it may not seem like I'm a fan of Juan Soto, that's quite the opposite. His combination of plate discipline, bat to ball skill, power, and overall presence is something that's truly rare and generational. There's no doubt in my mind Soto will remain a top player in baseball for the majority of his contract; however, I personally can't fathom paying $51 million a year to a glorified designated hitter. On the contrary, the following factors of; inflation, economics, marketability, and the opportunity to bring a World Series to Queens, will tell if this was a quality free agent deal.

It was extremely interesting to see the model rank Soto's contract as the second largest overpay for this time frame. Soto's 2022-204 seasons consisted of exceptional numbers that put him at the top of the league. However, the Mets gave him so much money, the model still predicts him as drastically overpaid.

This table shows the same idea as before, but for pitchers. This demonstrates the top 15 most overpaid contracts signed by pitchers from 2022-2025. While there's no outliers such as Soto in the last table, it's still interesting

	Player	Year	Actual AAV (M)	Predicted AAV (M)	Difference (M)
1	max scherzer	2022	43.33	16.64	-26.69
2	jacob degrom	2023	37.00	13.49	-23.51
3	justin verlander	2023	43.33	21.91	-21.42
4	robbie ray	2022	23.00	6.16	-16.84
5	walker buehler	2025	21.05	4.88	-16.17
6	noah syndergaard	2022	21.00	5.68	-15.32
7	luis severino	2025	22.33	10.12	-12.21
8	marcus stroman	2022	23.67	12.15	-11.52
9	lucas giolito	2024	19.25	8.71	-10.54
10	eduardo rodriguez	2024	20.00	9.92	-10.08
11	sean manaea	2025	22.01	12.62	-9.39
12	nathan eovaldi	2025	25.00	16.23	-8.77
13	nick martinez	2025	21.05	12.41	-8.64
14	martín pérez	2023	19.65	11.60	-8.05
15	matthew boyd	2025	14.50	6.53	-7.97

to see which contracts haven't worked out. One takeaway from this table is the majority of these deals are pitchers who are on the older side, or those who've struggled with injury issues. This may seem obvious even without this analysis, but it's still insightful to see these theories backed up.

Key Takeaways:

This project demonstrates the effectiveness of combining k-means clustering and regression modeling to predict free agent contracts in Major League Baseball. The most consistent predictors of AAV and contract length for both hitters and pitchers were WAR and Age. SLG and HR played a major role in statistical analysis for hitters, while vFA..pi. and ERA played a large role for pitchers. Clustering these players into different segment groups provided a critical improvement in model accuracy.

The "Very Good Hitters" and "Durable, Effective Starting Pitchers" clusters received the highest predicted and actual contracts, which aligns with basic baseball intuition. However, the model highlighted some inefficiencies in the speed/defense specialists, and power arms/volatility specialists. These cluster groups appeared to be slightly undervalued in actual signings, which suggests some teams may undervalue traits not directly related to offense of innings pitched.

The most interesting takeaway was that while linear models performed well, they consistently underpredicted the contract values of elite free agents. This was likely due to the linear structure of the model, and the absence of market/contextual variables such as; bidding wars, marketability, and agent influence.

Future Work:

This project was driven by strong personal interest and passion, and it will continue to evolve as efforts are made to enhance model accuracy, particularly for elite-level free agents. One key area for improvement involves exploring non-linear modeling techniques. Approaches such as piecewise regression, random forest, and XGBoost are well-suited to capturing the complex, non-linear relationships present in high-value contract negotiations. In addition to model structure, incorporating external variables will be critical. Factors like team needs, market size, positional scarcity, player marketability, and the effects of bidding wars all play a substantial role in determining contract outcomes and would increase the realism and predictive strength of the models. With these planned enhancements, the models are expected to produce more precise and representative forecasts for the most valuable free agents in future cycles.

Appendix:

Data Sources:

- Fangraphs.com - Player statistics (2022 through April 17, 2025).
- Spotrac.com - Free agent contract details, and upcoming free agent lists.
- Datasets created and imported into RStudio as CSVs;
  - ☐ hitters_clean_2022_2025.csv
  - ☐ pitchers_clean_2022_2025.csv
  - ☐ fa_2026_hitters.csv
  - ☐ fa_2026_pitchers.csv
  - ☐ hitters_fa_join
  - ☐ pitchers_fa_join

Data Dictionary:

- WAR: Wins Above Replacement.
- OBP: On Base Percentage.
- SLG: Slugging Percentage.
- HR: Home Runs.
- BB.: Walk Percentage.
- K.: Strikeout Percentage.
- SB: Stolen Bases.
- BsR: Baserunning Runs.
- Age: Player Age.
- ERA: Earned Run Average.
- xFIP: Expected Fielding Independent Pitching.
- FIP: Fielding Independent Pitching.
- K.9: Strikeouts per 9 Innings.
- BB.9: Walks per 9 Innings.
- IP: Innings Pitched.
- vFA..pi.: Average Fastball Velocity.
- AAV: Average Annual Value of Contract.
- Years: Length of Contract in Years.
- Cluster_Name: Categorical Label from K-means Output.
- PC1, PC2: Principal Components from PCA.

All Modeling specifications, variable imputations, and assumptions are fully detailed in the main body under each respective section.