Intro

Ever wonder how MLB hitters stack up beyond the usual "batting average" or "home runs" discussion? With baseball being an intensely data-driven sport, advanced statistics have become the norm for understanding players' strengths and weaknesses. In this analysis, I take a closer look at **350 MLB players** from the 2024 season, capturing a wide range of offensive metrics—like runs, hits, home runs, strikeouts, and more.

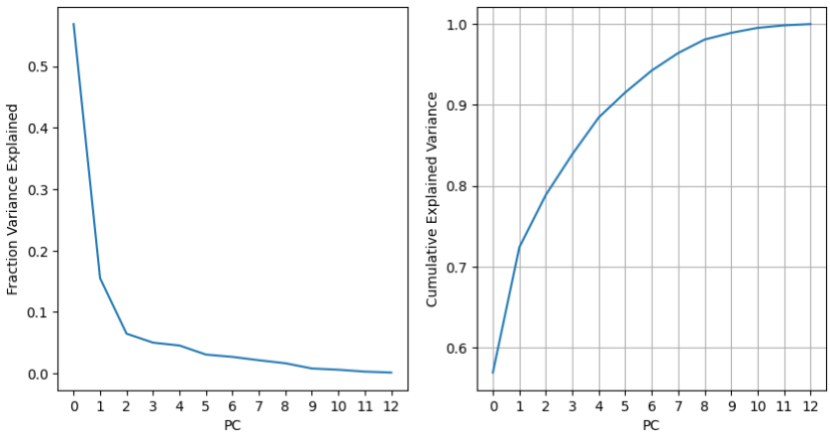
My main goals here are:

- **Dimensionality Reduction (PCA):** Identify the core factors that explain most of the variation among hitters.
- **Clustering (K-Means):** Group players into natural "archetypes," revealing similarities I might miss by just looking at batting average alone.
- **Classification (SVM):** Predict a player's batting performance category (e.g., over 0.325, or under 200, etc.) using the most influential stats.

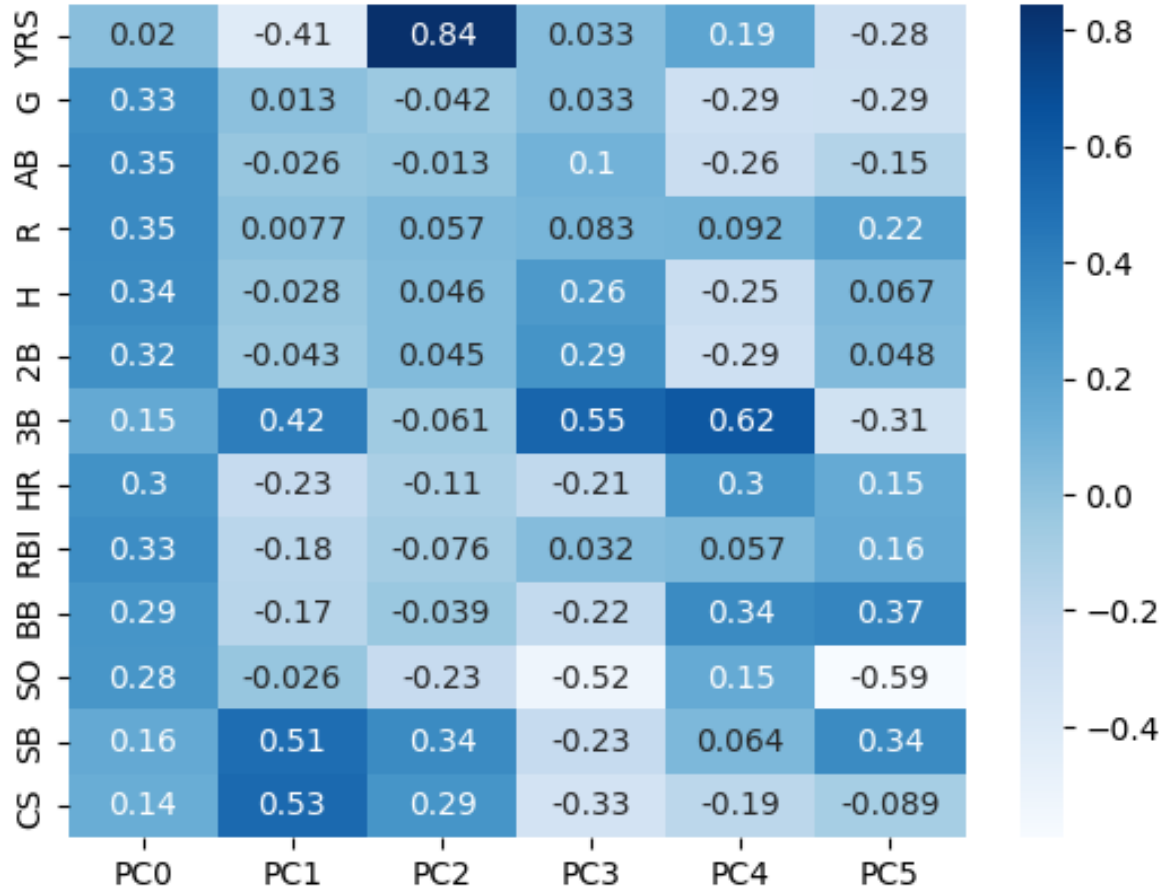
By the end of this post, you'll see which stats define a hitter's profile, how well these stats can group similar players together, and how accurately I can classify someone's hitting category based on their numbers. Whether you're a **fantasy baseball fanatic**, a **stat-savvy sportswriter**, or just a **curious fan**, this deep dive into batting data will give you a fresh perspective on what makes today's MLB hitters stand out—and how they compare to one another.

PCA and Key Dimensions

To cut through the noise of 13 different batting stats, I turned to PCA. Think of it as a mathematical spotlight that shines on the biggest differences among players.



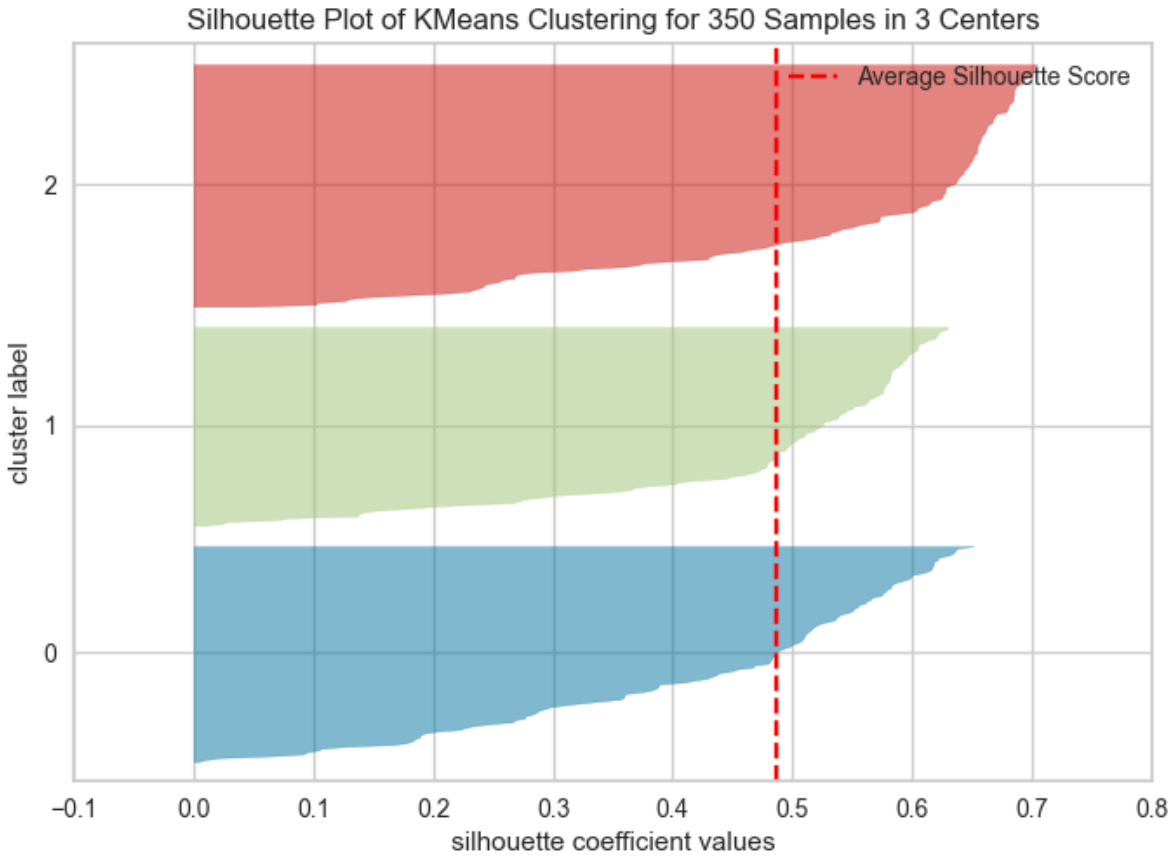
In our case, the first principal component (PC0) captured over half of the variability all by itself—a huge slice of the pie! This component seems strongly related to overall offensive volume: stats like **games played** (G), **at bats** (AB), **hits** (H), and **home runs** (HR) stand out.



The second principal component separated speedsters from sluggers, emphasizing stolen bases (SB) and triples (3B) on one side and power hitters (HR) on the other. A third, smaller component zeroed in on experience (YRS), highlighting how veterans differ statistically from rookies.

Clustering with K-Means

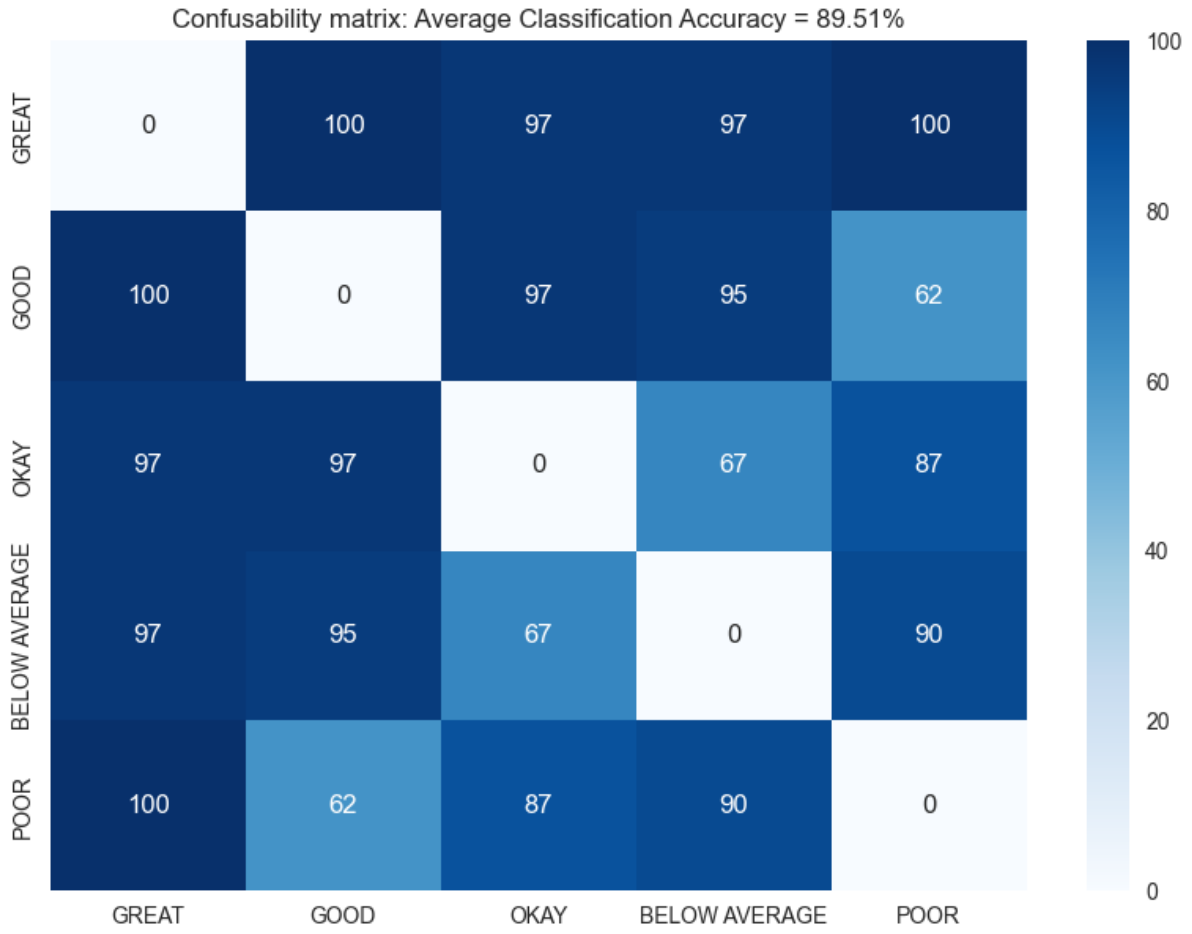
With 13 features in hand, I asked K-Means to form clusters—essentially letting the data group players with similar traits. I found that three clusters gave us a silhouette score of about 0.41, indicating a decent separation among groups.



One cluster leaned more toward our sluggers (high HR, RBI), another leaned toward speedy/contact hitters (more stolen bases and triples…Bobby Witt anyone?), and the last cluster showed a balance of both more representing MLB's all-arounders. I also tried 5 clusters, but that ended up with a lower silhouette score (~0.22), meaning more overlap—so I stuck with three clear-cut archetypes.

SVM Classification

Next, I tried to predict a player's batting average category—ranging from GREAT (BA ≥ .325) down to POOR (BA < .200)—using a linear SVM model. I tested one-on-one matchups between categories (e.g., GREAT vs. GOOD, GOOD vs. OKAY) and found that across all pairwise comparisons, I hit about **90% accuracy** on average. Not bad for a handful of stats like Runs (R), At Bats (AB), Hits (H), RBI, and Games (G). Interestingly, adding even more features only bumped accuracy a small bit, suggesting these five carry the most weight.



Reading the Confusability (Confusion) Matrix:

Darker blue cells here represent *higher* accuracy when trying to tell one category apart from another (e.g., "GREAT" vs. "GOOD"). Lighter blue cells mean the model had more trouble distinguishing between those two categories, leading to more mistakes (misclassification). So, whenever you see a lighter cell, that row's category is often mixed up with the column's category.

Insights and Next Steps

1. **A Single Dimension Explains ~50%**: Overall exposure in the lineup (games, at bats, etc.) is a massive driver in how hitters vary.
2. **Speed vs. Power**: Our second PCA dimension clearly teased out stolen-base threats vs. long-ball specialists.
3. **Three Distinct Clusters**: I identified three player archetypes that capture different playing styles, with moderate overlap.

4. **~90% Accuracy in Classification**: Using an SVM approach, I can fairly reliably categorize hitters by batting average bracket—though ‘OKAY’ vs. ‘BELOW AVERAGE’ was trickier.

While these models help us see how different batting attributes shape performance, there’s more to explore. Adding advanced metrics like exit velocity, launch angle, or sprint speed could make each cluster even more precise. I could also try other algorithms, from Random Forests to neural nets, to see if I can push classification accuracy higher. And for deeper insight, exploring a time-series approach (e.g., how a hitter changes year over year) might reveal truly unique career arcs.