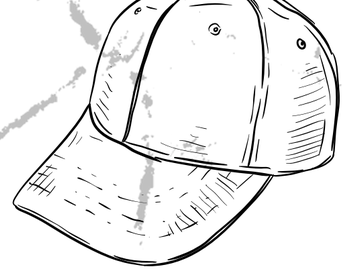
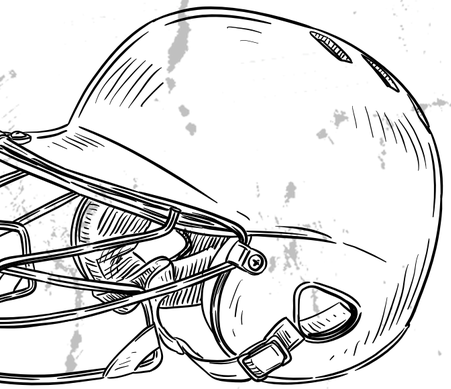
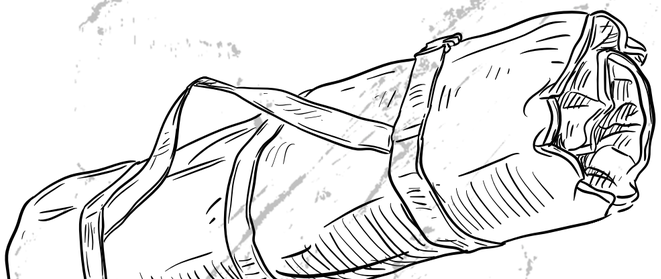


MLB OFFENSIVE ARCHETYPES: CLUSTERING & PERFORMANCE



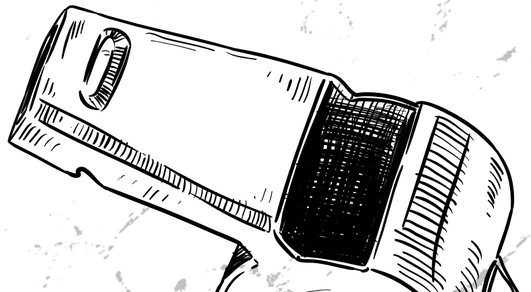
CADEN ZADELL | BSAN 360 FINAL PROJECT | 12/11/2025

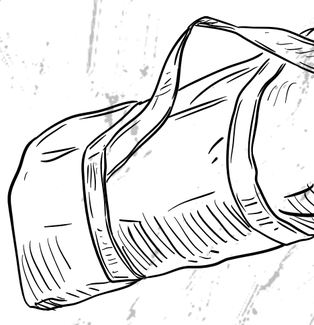
GITHUB

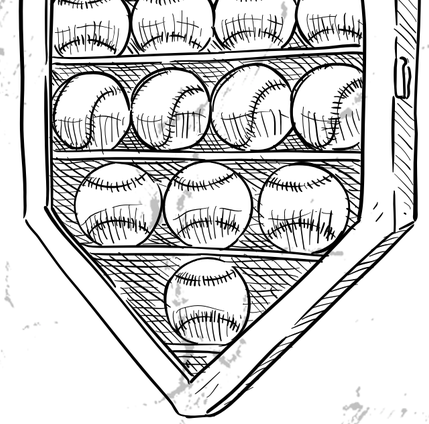


DATASET OVERVIEW

- **Dataset: Kaggle MLB Hitting & Pitching Stats**
  - Only used offensive stats
- Original Shape: (2508, 18)
- Cleaned Shape: (2496, 18)
  - Focus on batting statistics: **AVG, OBP, SLG, OPS, Hits, HR, RBIs, Runs.**
  - Dataset spans thousands of MLB seasons, giving >45,000 player-season observations.
  - **Goal:** identify offensive performance patterns and hitter archetypes.





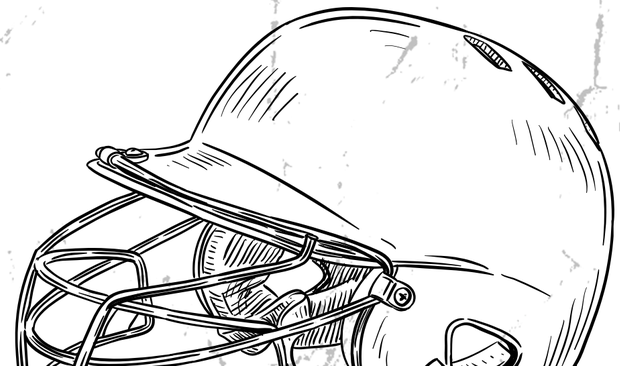


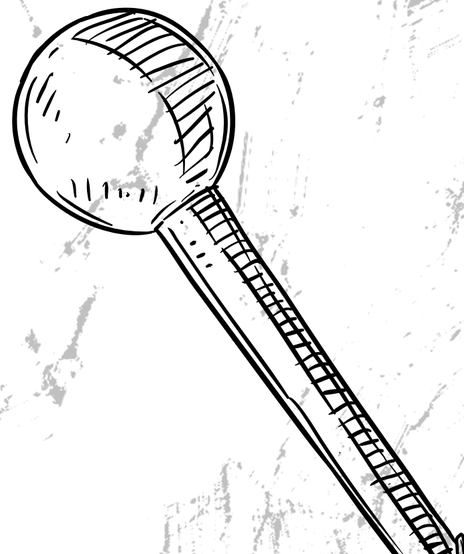
RESEARCH QUESTIONS

1. What is the overall offensive performance distribution of MLB hitters? How does batting average relate?

2. How does offensive performance differ by defensive position?

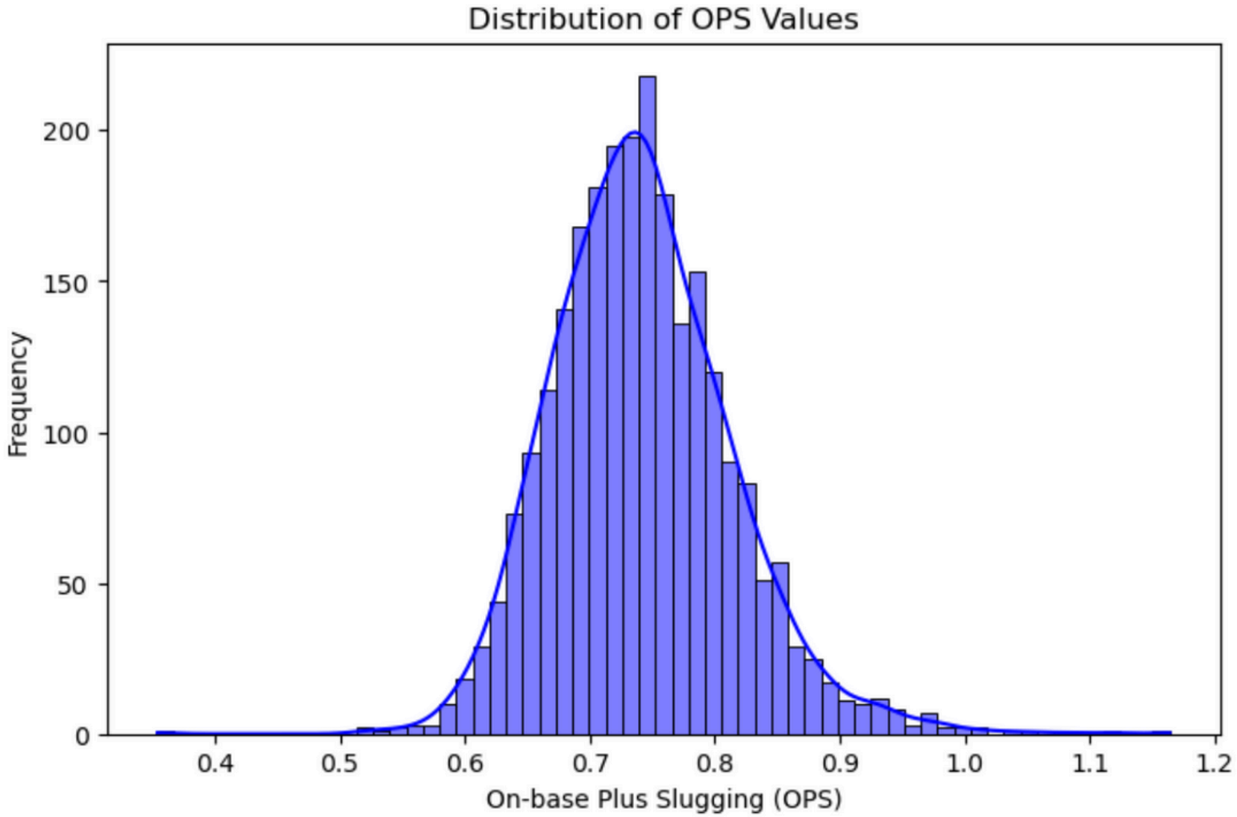
3. Do hitters naturally cluster into meaningful performance archetypes?





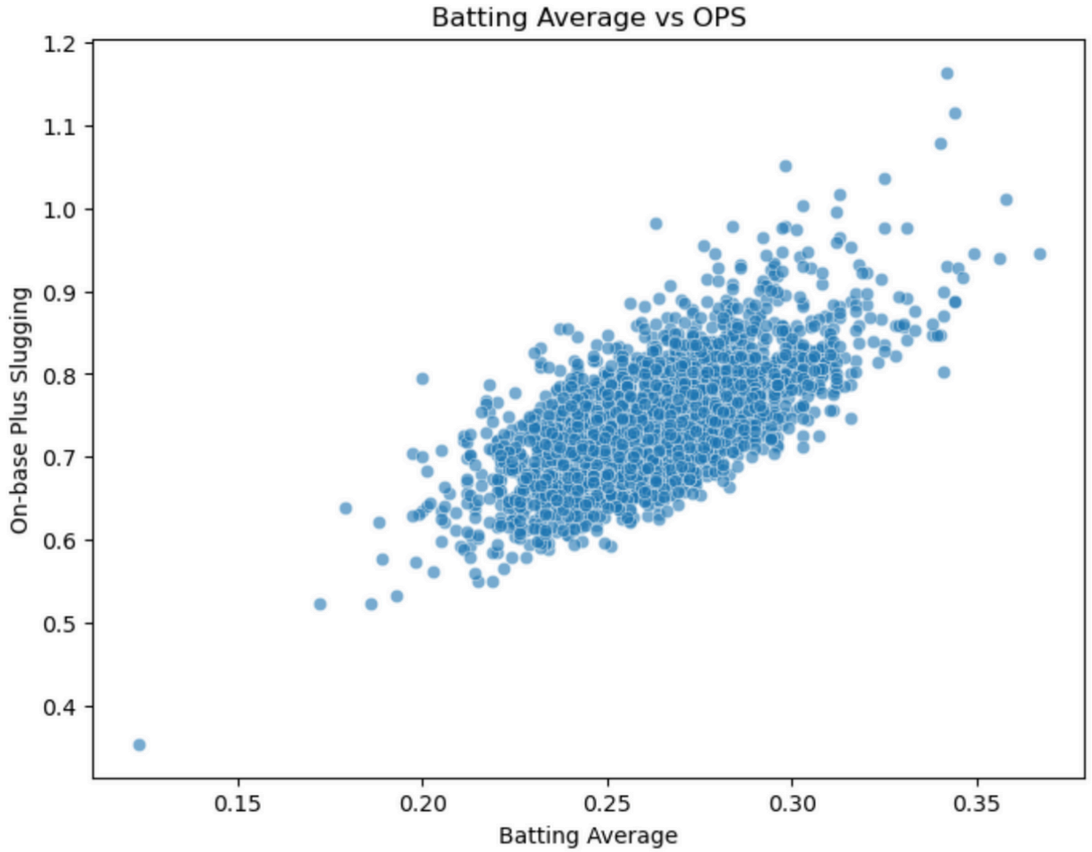
OFFENSIVE PERFORMANCE DISTRIBUTION

- OPS distribution is approximately normal with a peak around ~0.74
- Indicates most MLB hitters fall within a narrow offensive band
- Relevant because OPS is used as our primary performance metric
- Helps justify segmentation: hitters vary enough to form clusters



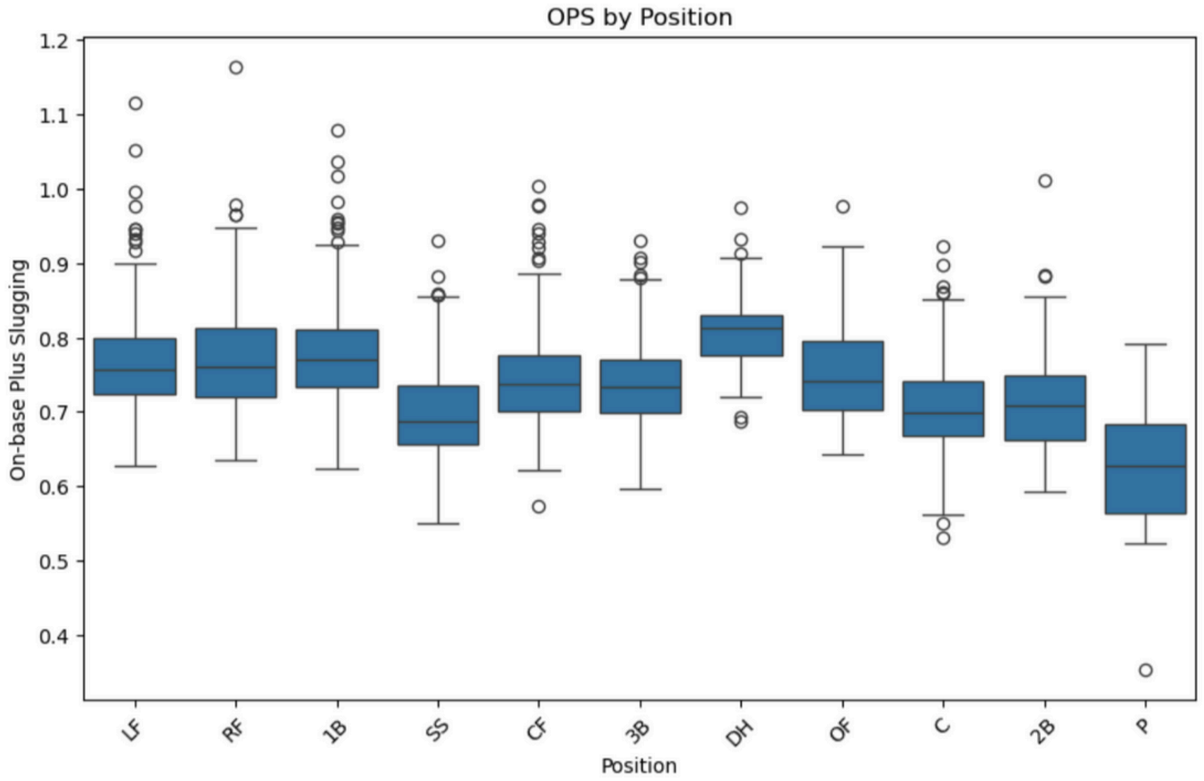
CORRELATION OF BATTING AVERAGE TO OPS

- Strong positive correlation: as Batting Average increases, OPS increases
- Suggests contact skill meaningfully contributes to overall offensive value
- Scatter plot shows consistent upward pattern across all hitters
- Useful in understanding hitter attributes before clustering



OFFENSIVE PERFORMANCE BY POSITION

- DH, LF, RF show highest median OPS — typical power positions
- SS, 2B, C display lower OPS due to defensive value emphasis
- Pitchers have lowest OPS, validating expected positional trends
- This informs team roster construction & salary valuation



OPS TIER CLASSIFICATION

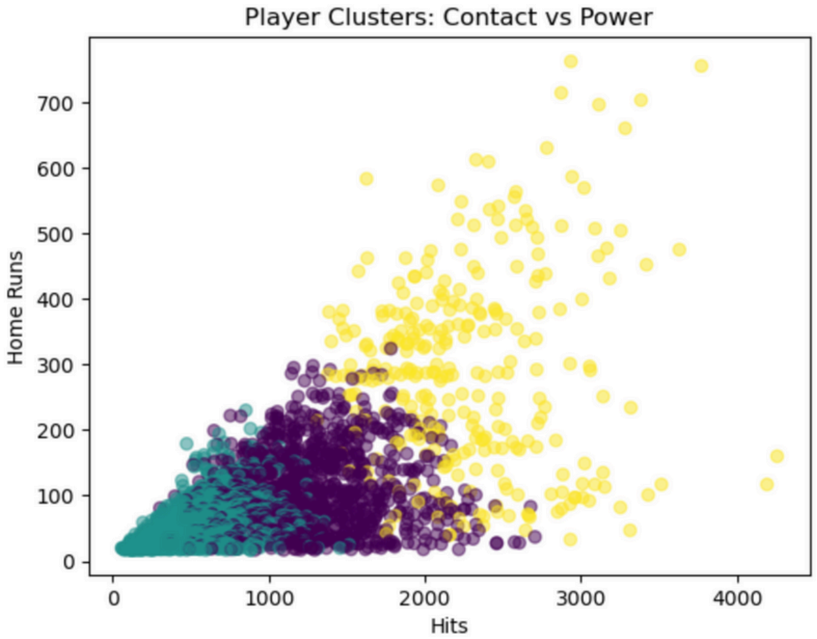
- Below Avg (<.700): 714 players
- Above Avg (.700–.800): 1320 players
- Great (.800–.900): 402 players
- Elite (>.900): 60 players
- Tiers help frame cluster interpretation and rarity of elite hitters.

```
print("\nNumber of players in each OPS tier:")
print(tier_counts)
```

```
Number of players in each OPS tier:
OPS_tier
Below Avg (<.700)          714
Above Avg (.700–.800)     1320
Great (.800–.900)          402
Elite (>.900)               60
Name: Player name, dtype: int64
```

DO HITTERS FORM NATURAL PERFORMANCE CLUSTERS?

- K-Means produced 3 clear clusters:
  - **Cluster 0:** Low-impact hitters (low HR, low Hits)
  - **Cluster 1:** Balanced hitters with solid but unspectacular power
  - **Cluster 2:** High-power hitters with extreme HR totals
- Visualization shows clear separation based on Hits and HRs.



Cluster	AVG	On-base Percentage	Slugging Percentage	Home Runs	RBIs	Runs	Hits	Steals
0	0.28	0.34	0.42	106.75	592.80	655.40	1276.82	105.94
1	0.25	0.31	0.39	52.21	244.40	247.39	513.53	26.40
2	0.29	0.37	0.48	289.02	1242.66	1256.08	2272.49	188.78

CONCLUSION & INSIGHTS

- Offensive performance follows a consistent distribution with meaningful variation.
    - OPS **strongly correlates** with Batting Average, validating OPS as a composite metric
    - Clustering reveals three hitter archetypes: **Low-impact, Balanced, and Power**
    - Defensive positions show expected offensive differences (DH > C/SS/P)
    - Findings can help MLB teams in **scouting, roster building,** and **contract evaluation**