# Descriptive Analysis Report – MLB Free-Agent Batters (1998–2013)

## 1 | Objective

The front office of a small-market MLB team requested a data-driven analysis of free-agent batters to identify salary trends, key value drivers, and potential bargains. This report evaluates 2,400+ free-agent seasons (1998–2013), using inflation-adjusted contract values to ensure comparability across years.

## 2 | Methodology

Data & Tools:

- Source: baseballcase_data.csv, cleaned to standardize headers.
- Key Fields: SalaryAdj (inflation-adjusted salary), Year, HR (home runs), OBP (on-base percentage), SLG (slugging), WAR (wins above replacement).

Processing:

- Removed rows with missing Salary, Year, or HR values.
- Converted numeric columns and computed:
  - log_salary: Log-transformed salary to address skew.
  - cost_per_HR: SalaryAdj ÷ HR (metric for HR efficiency).

Analytics:

- Descriptive statistics.
- Pearson correlations with salary.
- Visualizations:
  1. Salary distribution.
  2. Salaries by signing year.
  3. Correlation heatmap.
  4. Salary vs. HR scatterplot, highlighting bargains and outliers.
- Outputs: All visuals and tables saved for reference.

## 3 | Results & Visuals

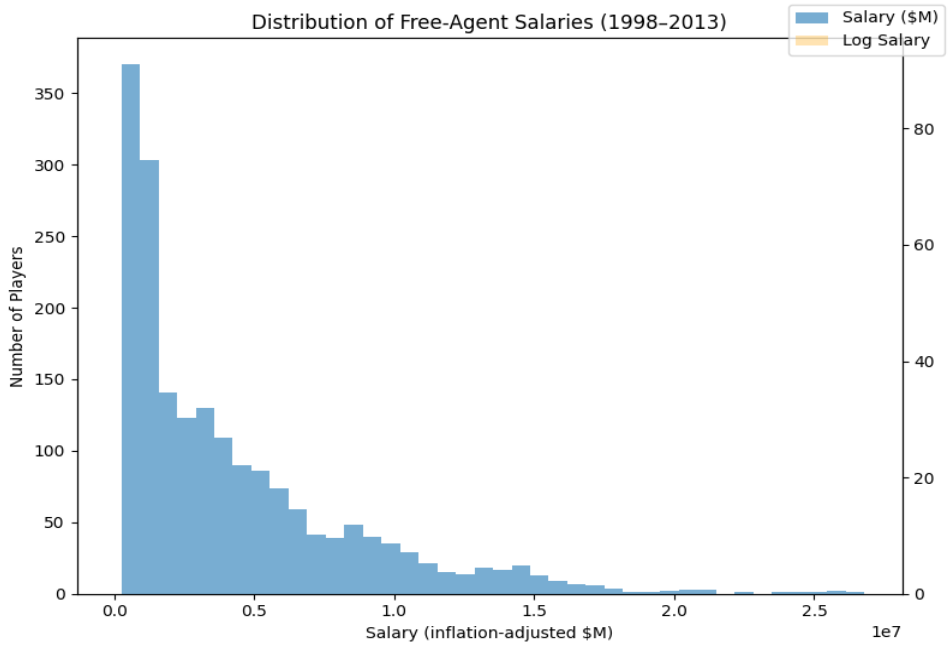


**Figure 1 – Salary Distribution**

- Right skewed: Over 50% of free agents signed below 5M,withalongtailexceeding5M,withalongtailexceeding25M.
- Log-salary transformation normalizes the distribution, validating its use in regression models.

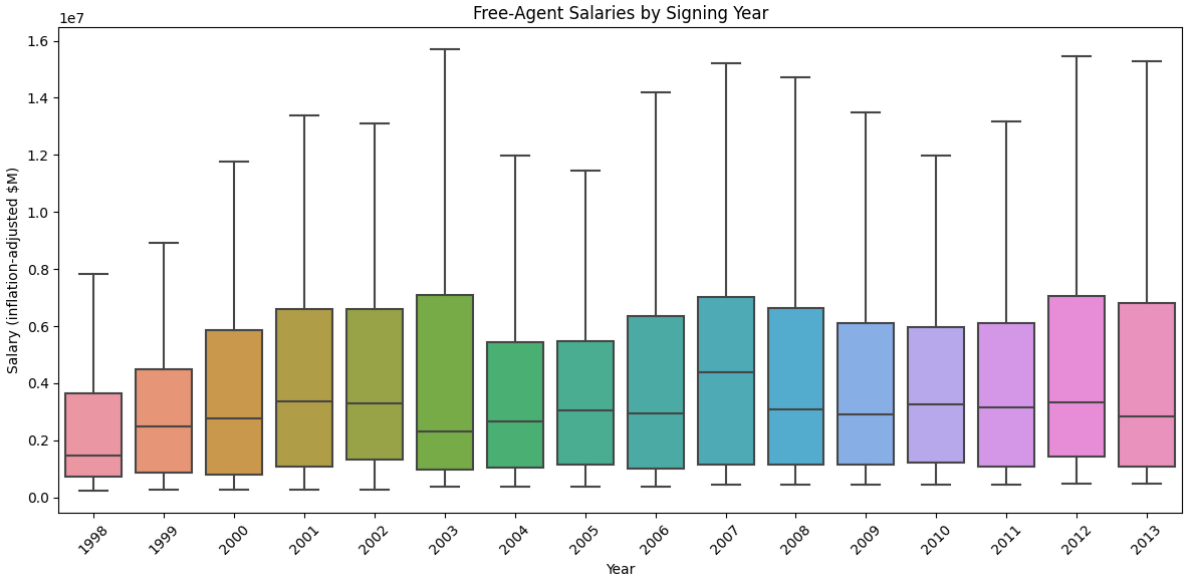


**Figure 2 – Salaries by Signing Year**

- Median salaries doubled: From 1.5M(1998)to1.5M(1998)to3M (2013).
- Dispersion widened: After 2005, the gap between superstar and mid-tier contracts grew sharply.

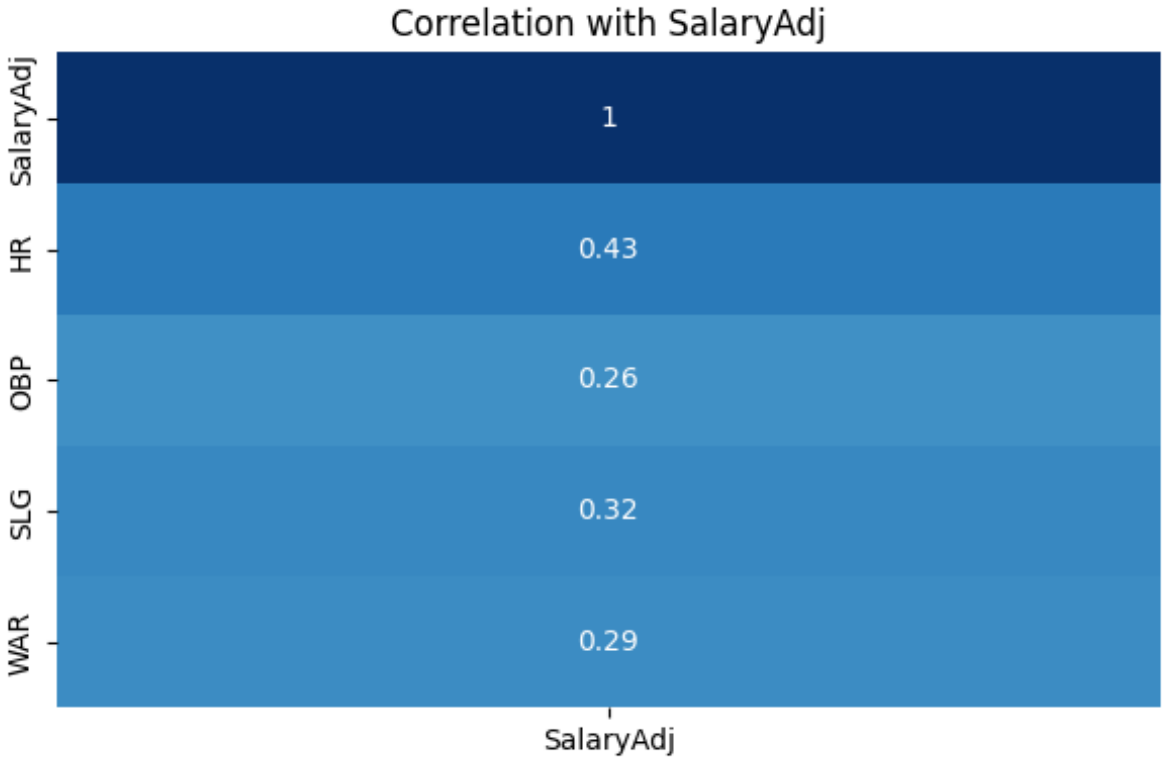


**Figure 3 – Correlation with Salary**

| | SalaryAdj |
|---|---|
| SalaryAdj | 1 |
| HR | 0.432 |
| OBP | 0.259 |
| SLG | 0.322 |
| WAR | 0.291 |

**Pearson R between SalaryAdj and HR / OBP / SLG / WAR**

- Strongest driver: Home runs (HR, r ≈ 0.43).
- Moderate drivers: Slugging (SLG, r = 0.32), WAR (r = 0.29).
- Weakest driver: On-base percentage (OBP, r = 0.26), suggesting power metrics are prioritized.

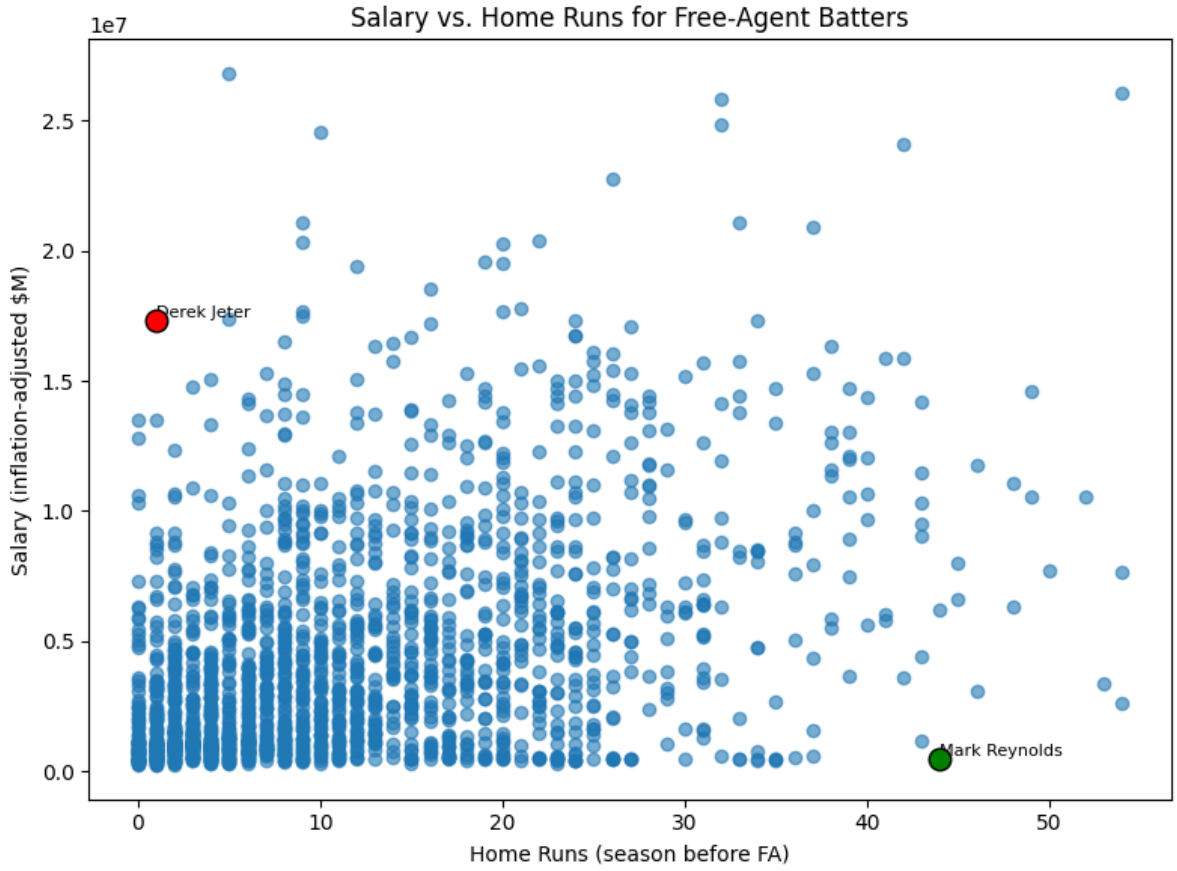


**Figure 4 – Salary vs. Home Runs**

| Player | HR | SalaryAdj | cost_per_HR |
|---|---|---|---|
| Luis Castillo | 1 | 6594891.15 | 6594891.15 |
| Marlon Byrd | 1 | 6697385.75 | 6697385.75 |
| Lance Johnson | 1 | 7274772.56 | 7274772.56 |
| Rey Ordonez | 1 | 8240823.14 | 8240823.14 |
| John Valentin | 1 | 8509933.3 | 8509933.3 |
| Tony Gwynn | 1 | 8718935.2 | 8718935.2 |
| Jeff Cirillo | 1 | 8795623.4 | 8795623.4 |
| Mark Kotsay | 1 | 9175500.73 | 9175500.73 |
| Jason Kendall | 1 | 13521285.6 | 13521285.6 |

| | | | |
|---|---|---|---|
| Derek Jeter | 1 | 17331985.8 | 17331985.8 |

TOP 10 cost_per_HR

**TOP 10**

| Player | HR | SalaryAdj | cost_per_HR |
|---|---|---|---|
| Mark Reynolds | 44 | 469719.19 | 10675.4361 |
| Aramis Ramirez | 34 | 381941.888 | 11233.5849 |
| Nick Swisher | 35 | 394139.553 | 11261.1301 |
| Vernon Wells | 23 | 291395.506 | 12669.3698 |
| Carlos Gonzalez | 34 | 441501.186 | 12985.329 |
| Adam Lind | 35 | 457823.343 | 13080.6669 |
| Jack Cust | 33 | 452511.732 | 13712.4767 |
| Bill Hall | 35 | 491792.04 | 14051.2011 |
| Paul Goldschmidt | 36 | 509764.289 | 14160.1192 |
| Hanley Ramirez | 33 | 484518.659 | 14682.3836 |

BOTTOM 10 cost_per_HR

**BOTTOM 10**

- Cluster: Most players fall below 25 HR and $10M.
- Outliers:
  - Mark Reynolds (2009): 44 HR at ~0.8M(cost:<0.8M(cost:<20k per HR).
  - Derek Jeter (2011): $17M salary despite 0 HR (intangibles/brand premium).

**4 | Key Insights**

1. Salary inflation & inequality: Median salaries doubled, but top-tier contracts grew faster, reflecting a star-driven market.
2. Power pays: HR remains the strongest salary predictor; OBP's influence has plateaued post-Moneyball.
3. WAR undervalued? Teams recognize overall value (WAR correlation = 0.29) but prioritize HRs.
4. Bargain targets: Players like Mark Reynolds (2009) and José Bautista (2010) delivered 25+ HR at <$2M.
5. Overpay risks: Aging stars (e.g., Jeter) or low-power hitters often command inflated salaries.

**5 | Recommendations**

- Target undervalued power: Prioritize mid-tier sluggers (25–35 HR) with salaries below $5M.
- Model log-salary: Use log-transformed salary for predictive analytics to handle skew.
- Expand variables: Incorporate leadership, marketability, and postseason performance to explain high-salary outliers.
- Monitor trends: Post-2013 data may reveal new salary plateaus; update analysis annually.