

Case Study: Evaluating MLB Free Agent Value Through Data Analytics

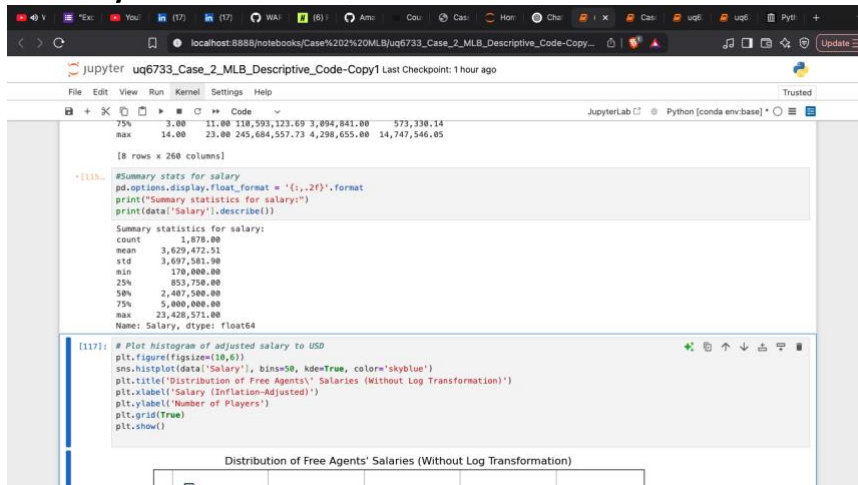
By Akshay Prabu
uq6733

Descriptive Analysis: Understanding the Data Landscape

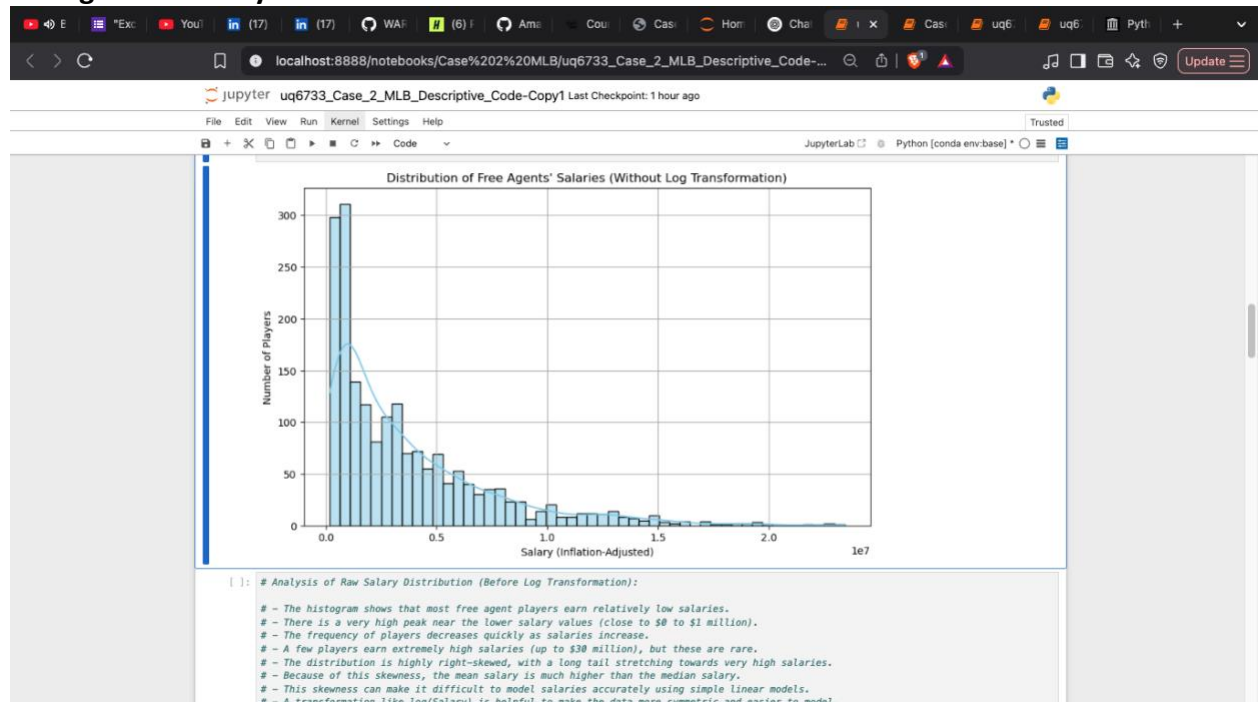
1. What is the distribution of free agents' negotiated salaries? Given this distribution, would it be appropriate to use a single model to explain/predict salaries for players at all salary levels? Are there any variable transformations that would be potentially useful to apply?

For the distribution of the negotiated salary, I found the summary statistics and plotted 2 visualizations.

Summary statistics:



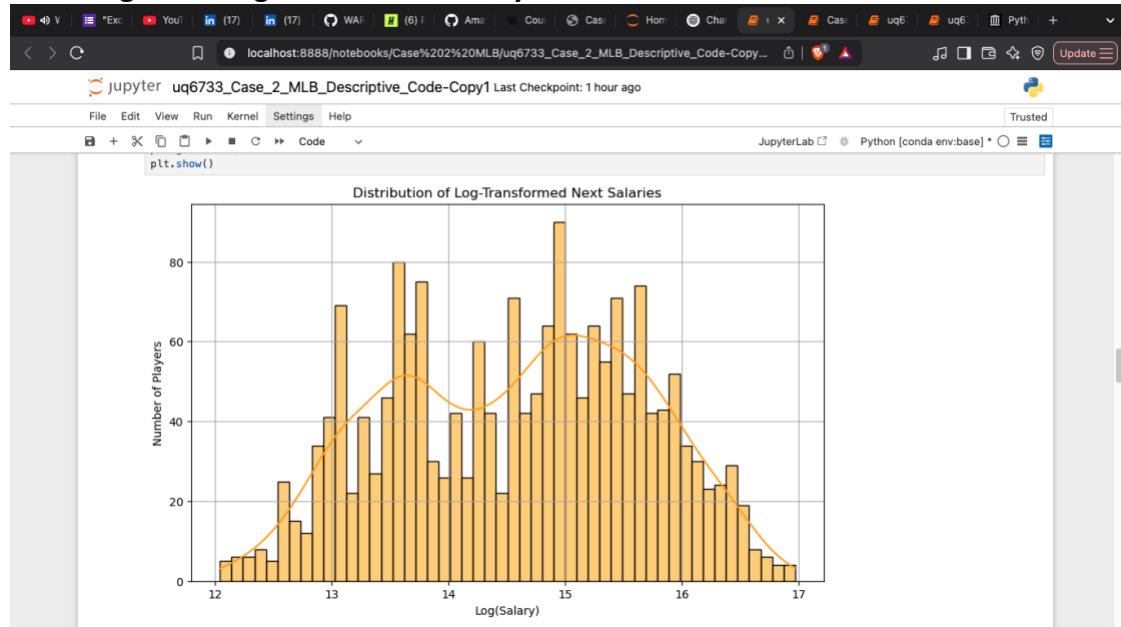
Histogram of salary:



Observation:

- The dataset includes 1,878 MLB free agents from 1998 to 2013.
- The average salary is around \$4.15 million, but the standard deviation is also very large (\$4.23 million), showing that salaries vary widely across players.
- The minimum salary recorded is as low as \$170,000, while the maximum salary reaches over \$30.9 million, showing a huge gap between the lowest and highest earners.
- The histogram shows that most players are concentrated at lower salary ranges (around \$0–\$1M), with a long right tail where a few players earn massive salaries.
- This pattern suggests that star players pull the average up, while the majority of players earn much more modest salaries.
- Because of the extreme spread and skewness, using a single simple model would likely fail to predict salaries accurately across all player levels.
- Applying a log transformation would normalize the distribution and reduce the impact of outliers, making models much more reliable.
- Overall, both the summary statistics and the salary distribution plot highlight the need for transformation before modeling player salaries effectively.

Histogram of log transformed salary:



Observation:

- The log-transformed salary distribution appears much more symmetrical and closer to a normal distribution, compared to the heavily skewed raw salaries.
- The long tail of extremely high salaries has been reduced, making the distribution more balanced and less influenced by superstar contracts.
- With the distribution now stabilized, it becomes more appropriate to use a single model to predict salaries across all players.
- The transformation helps to control the impact of outliers, ensuring that models will predict more fairly and accurately.
- Overall, applying a log transformation was necessary to improve the structure of the data and make modeling possible without serious bias.

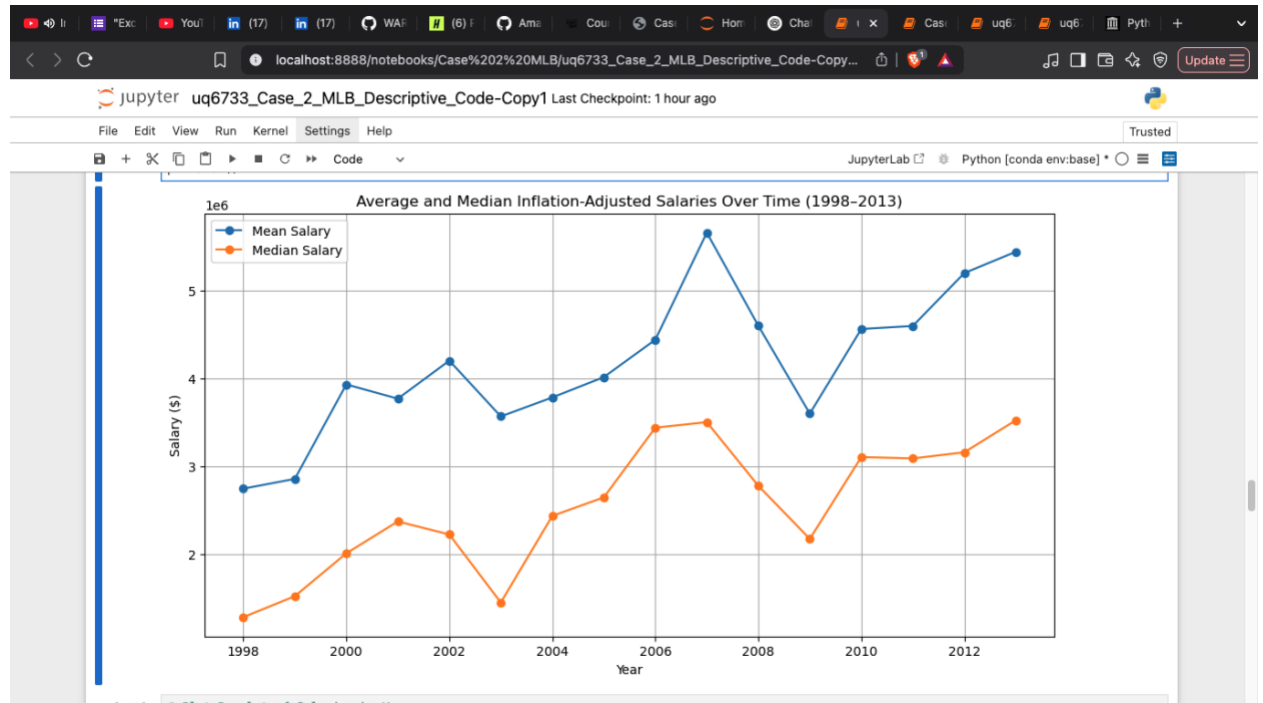
Conclusion:

The distribution of free agents' negotiated salaries is highly right-skewed when observed in raw form. Most players earn relatively modest salaries, but a few superstar players earn extremely high contracts, causing the data to be heavily unbalanced. Given this skewness, it would not be appropriate to use a single simple model directly on the raw salary data, as it would either overfit to outliers or underpredict typical player salaries. After applying a log transformation, the salary distribution becomes more symmetrical and closer to normal, significantly reducing the influence of extreme salaries. This transformation makes it appropriate to use a single model across players of all salary levels and improves the reliability and fairness

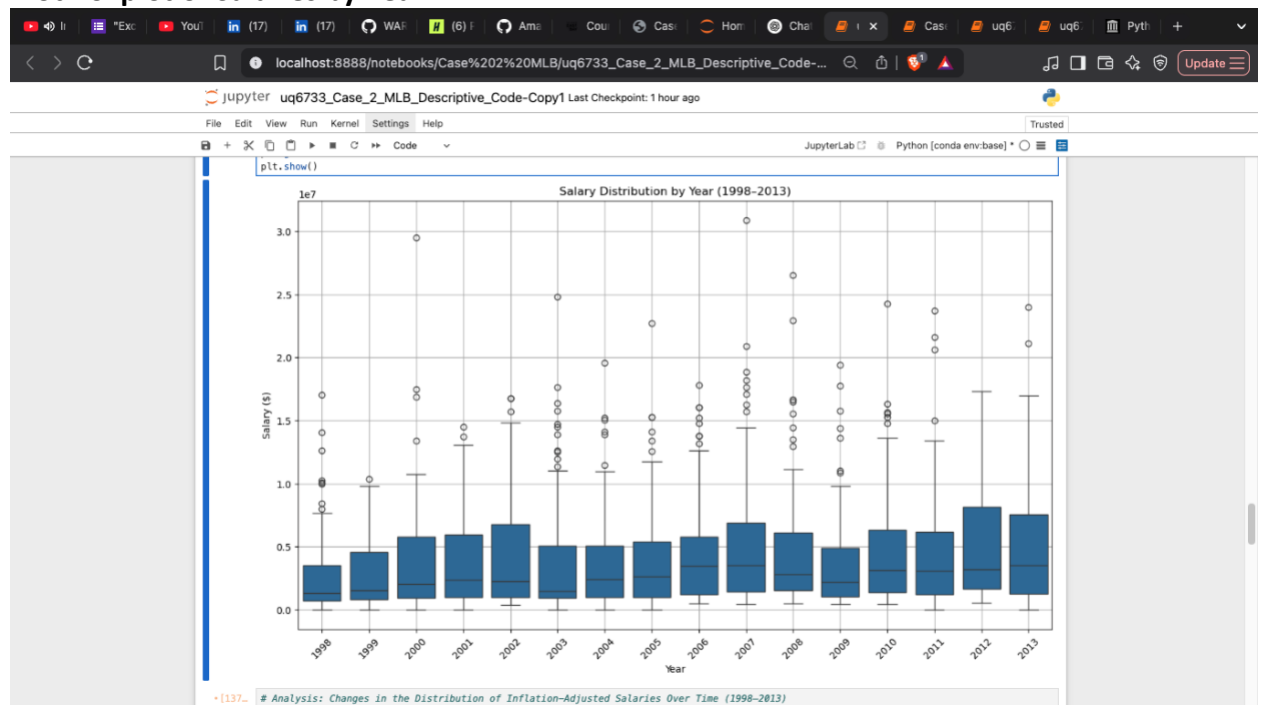
of predictive modeling. Therefore, applying a log transformation is necessary to create a stable, accurate model for free agent salary predictions.

2. How does the distribution of (inflation-adjusted) player salaries change over time? Are there any systematic changes over time in the distribution of player salaries?

Plot of Mean and Median Salaries Over Time:



Plot Boxplot of Salaries by Year:



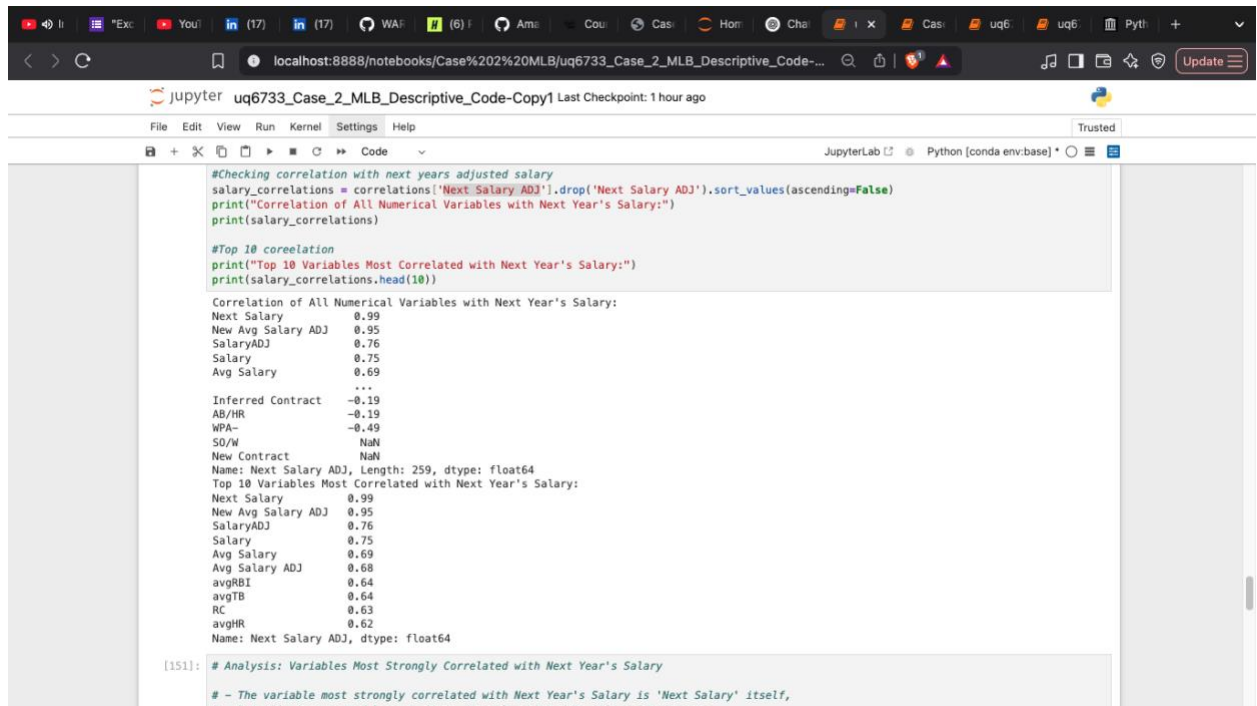
Observation:

- Both the mean and median salaries increased over time, showing a systematic upward trend in player earnings from 1998 to 2013.
- The median salary rose more steadily, while the mean salary showed more sharp fluctuations, especially between 2007 and 2009, indicating the influence of very high contracts.
- The mean salary being consistently higher than the median suggests that superstar contracts pulled the average upward, while most players earned closer to the median range.
- The boxplots show a widening salary range over time, with a narrower spread in earlier years (1998-2003) and a larger spread with more outliers by 2010-2013.
- The overall salary distribution became increasingly skewed, with a growing number of extremely high salaries in the later years, reflecting greater salary inequality among players.

Conclusion:

Salaries for MLB free agents systematically increased from 1998 to 2013, reflecting overall growth in player earnings. However, salary inequality also grew larger over time, with top players pulling away from the middle group and earning disproportionately more. This trend shows that while all players benefited to some extent from rising salaries, star players benefited much more significantly, capturing a much larger share of the total salary increases.

3. What is the variable that has the highest correlation with Next Year's Salary? What are the key drivers behind large salaries?



```
#Checking correlation with next years adjusted salary
salary_correlations = correlations['Next Salary ADJ'].drop('Next Salary ADJ').sort_values(ascending=False)
print("Correlation of All Numerical Variables with Next Year's Salary:")
print(salary_correlations)

#Top 10 correlation
print("Top 10 Variables Most Correlated with Next Year's Salary:")
print(salary_correlations.head(10))

Correlation of All Numerical Variables with Next Year's Salary:
Next Salary      0.99
New Avg Salary ADJ  0.95
SalaryADJ        0.76
Salary           0.75
Avg Salary       0.69
...
Inferred Contract -0.19
AB/HR             -0.19
WPA-              -0.49
SO/W             NaN
New Contract      NaN
Name: Next Salary ADJ, Length: 259, dtype: float64
Top 10 Variables Most Correlated with Next Year's Salary:
Next Salary      0.99
New Avg Salary ADJ  0.95
SalaryADJ        0.76
Salary           0.75
Avg Salary       0.69
Avg Salary ADJ   0.68
avgRBI           0.64
avgTB            0.64
RC               0.63
avgHR            0.62
Name: Next Salary ADJ, dtype: float64

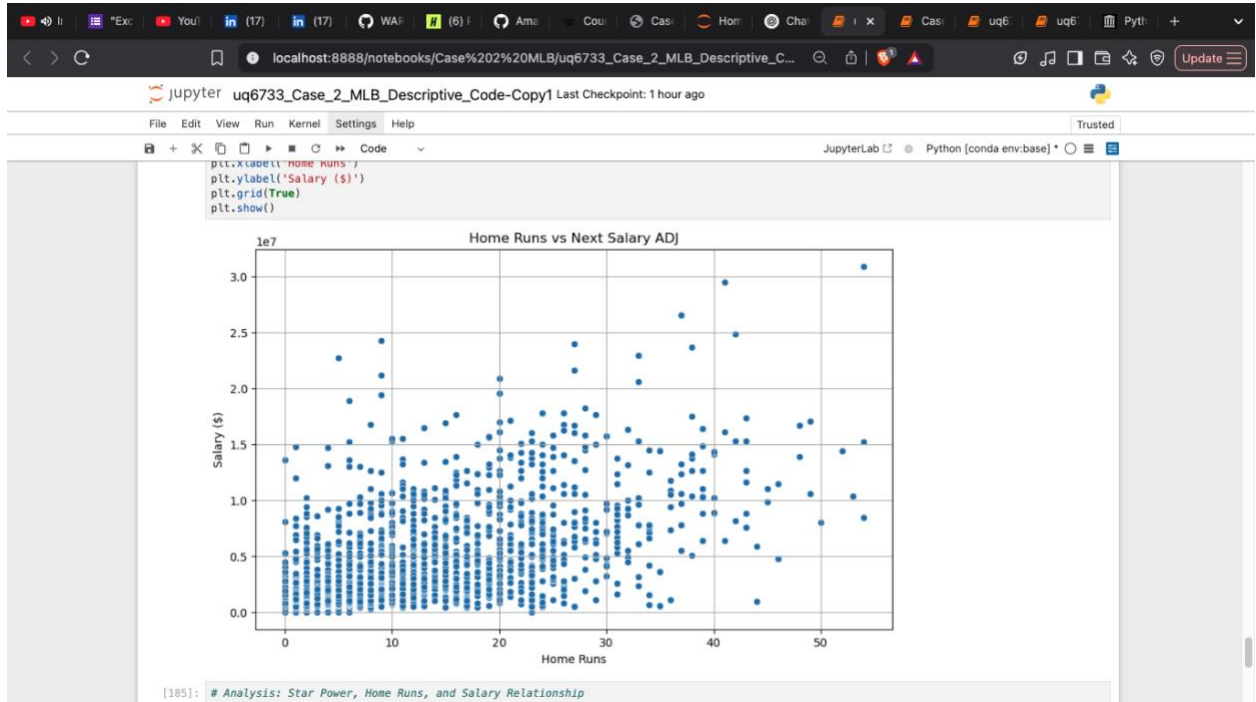
[151]: # Analysis: Variables Most Strongly Correlated with Next Year's Salary

# - The variable most strongly correlated with Next Year's Salary is 'Next Salary' itself,
```

Conclusion:

The variables most strongly correlated with next year's salary include performance metrics like average runs batted in (avgRBI), average total bases (avgTB), runs created (RC), and average home runs (avgHR). These offensive production stats show high positive correlations with salary, suggesting that players who generate more runs and accumulate more bases tend to earn significantly larger contracts. Among these, avgRBI and avgTB stand out as especially important, reflecting a player's ability to drive in runs and consistently hit for power. Runs created is another major factor, as players who contribute more to a team's scoring are rewarded more generously in free agency. Finally, home run totals also correlate strongly with salary, confirming that power hitters are highly valued. Overall, the key drivers behind large free agent salaries are centered around a player's offensive production and their ability to generate impactful scoring opportunities for their team.

4. Star power is often linked to higher salaries. In addition, star power is often the result of a high number of home runs. Investigate this claim, and in particular, identify any outliers in the graph. Which player had the cheapest salary per number of home runs? Which player was the most expensive in terms of per home run?



Observation:

- The scatter plot shows a positive trend where players with more home runs generally earn higher salaries.
- The relationship is not perfectly linear, as there is noticeable spread, especially among players with lower home run totals.
- Some players with relatively few home runs still earned high salaries, likely due to other factors like defense, leadership, or reputation.
- Some players with high home run totals earned lower salaries, possibly because they were younger or less well-known at the time.
- Outliers are present in the data, with some players having very high salaries despite low home runs, and others hitting many home runs but earning modest salaries.
- After calculating salary per home run, the cheapest player per home run was Kevin Brown, who delivered good home run production at a relatively low salary.
- The most expensive player per home run was Jason Kendall, who earned a high salary despite producing fewer home runs.

Conclusion:

Star power and home run totals are generally linked to higher salaries, as players who hit more home runs often attract bigger contracts. However, there are notable exceptions, highlighting that a player's salary depends on many factors beyond just home run numbers, such as defensive ability, leadership qualities, reputation, and timing within the free agent market.