Data Wrangling ASG 2

Tan Jun Yu Zavier
(S10255651D)

Table of contents

01
Introduction

01-Introduction

1871 - 2014

AIM

How might we Predict MLB players' salaries using performance metrics

Role

Data Analyst

Outcome

Provide valuable insights to potential management team, optimising allocation of financial resources and making strategic decisions.

Datasets

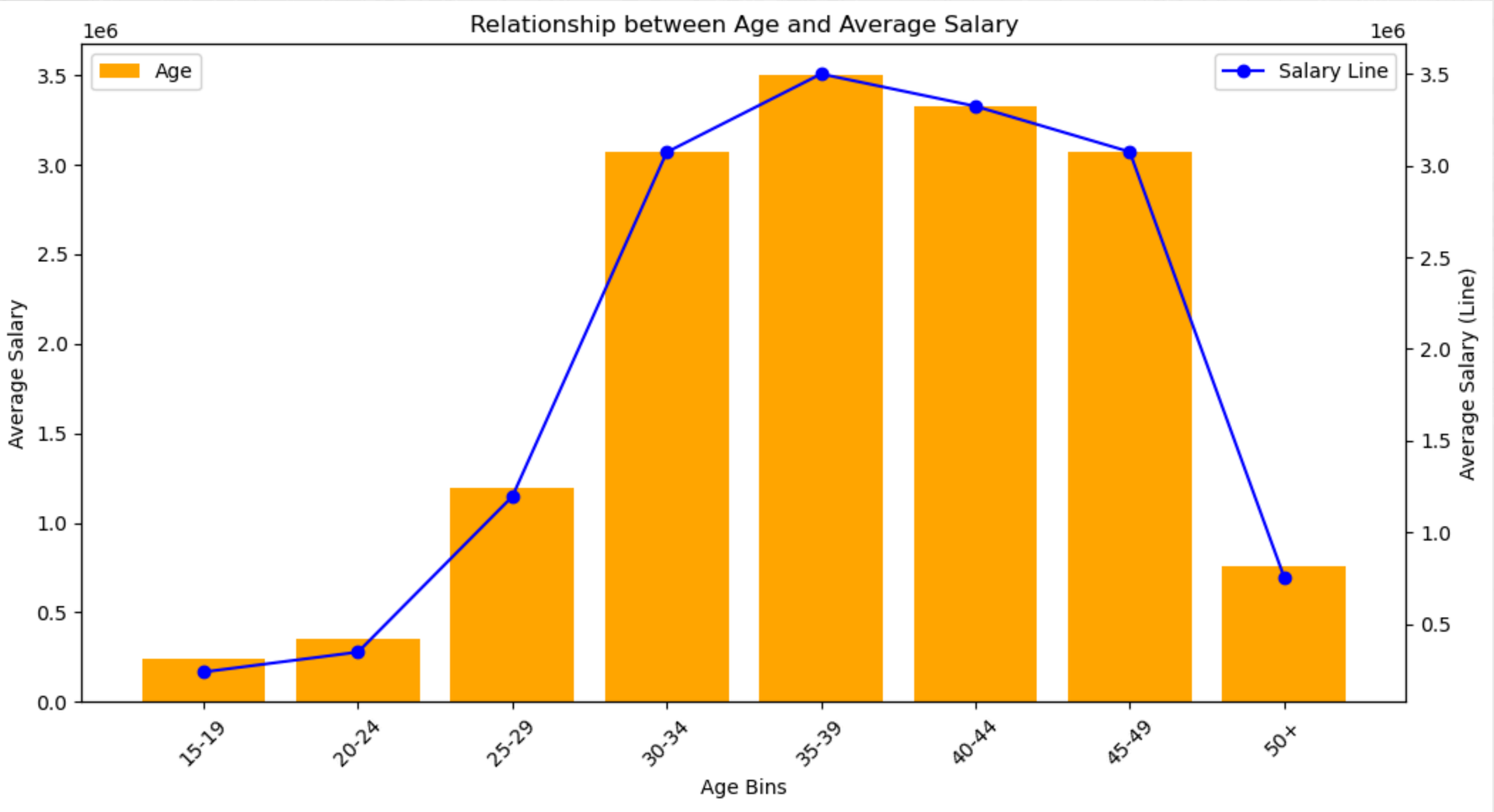
Datasets used are

1. salary
2. player
3. fielding
4. pitching
5. batting

01-Target Variable

02 Data Exploration

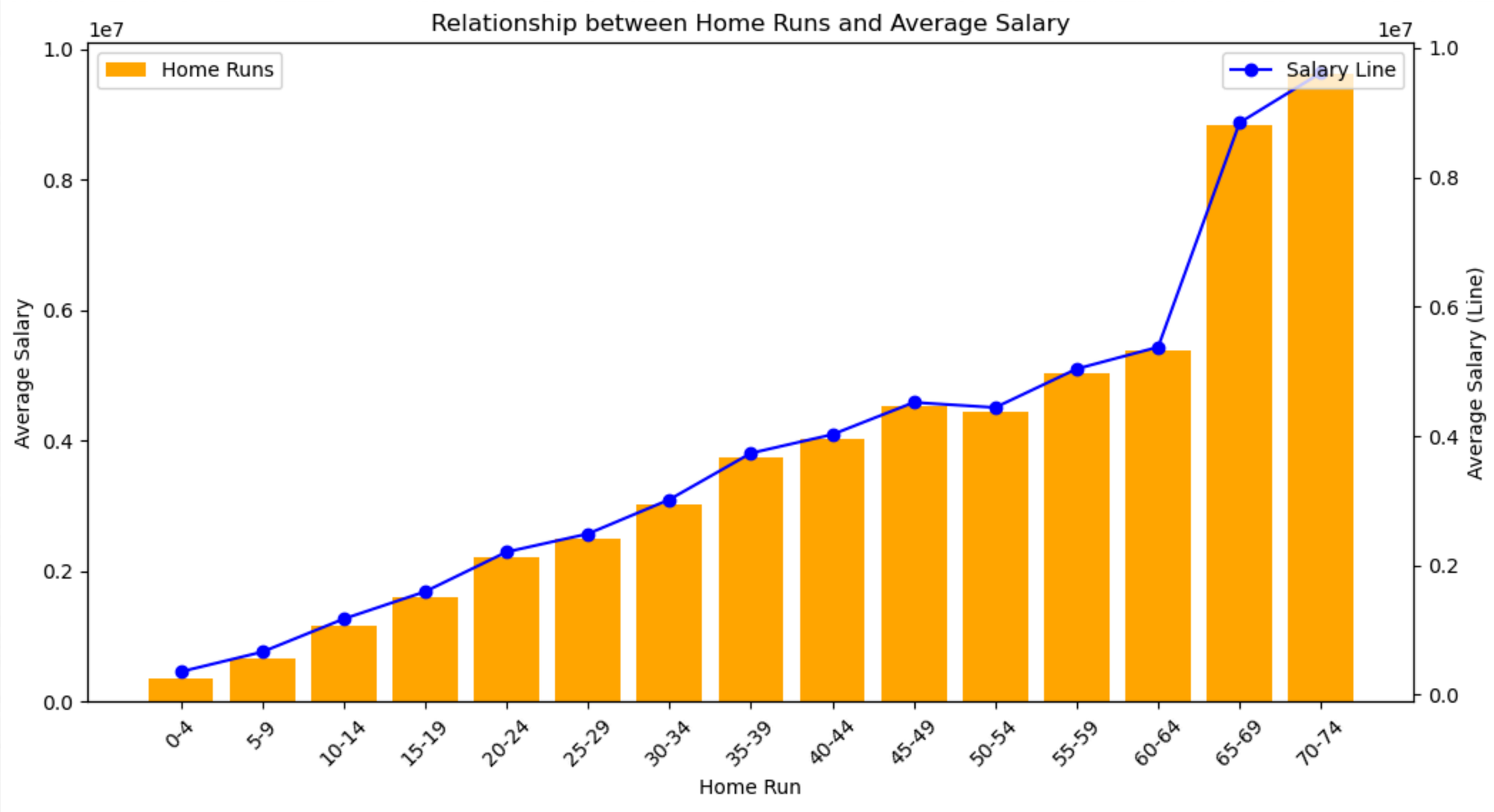
Salary & Age Relationship



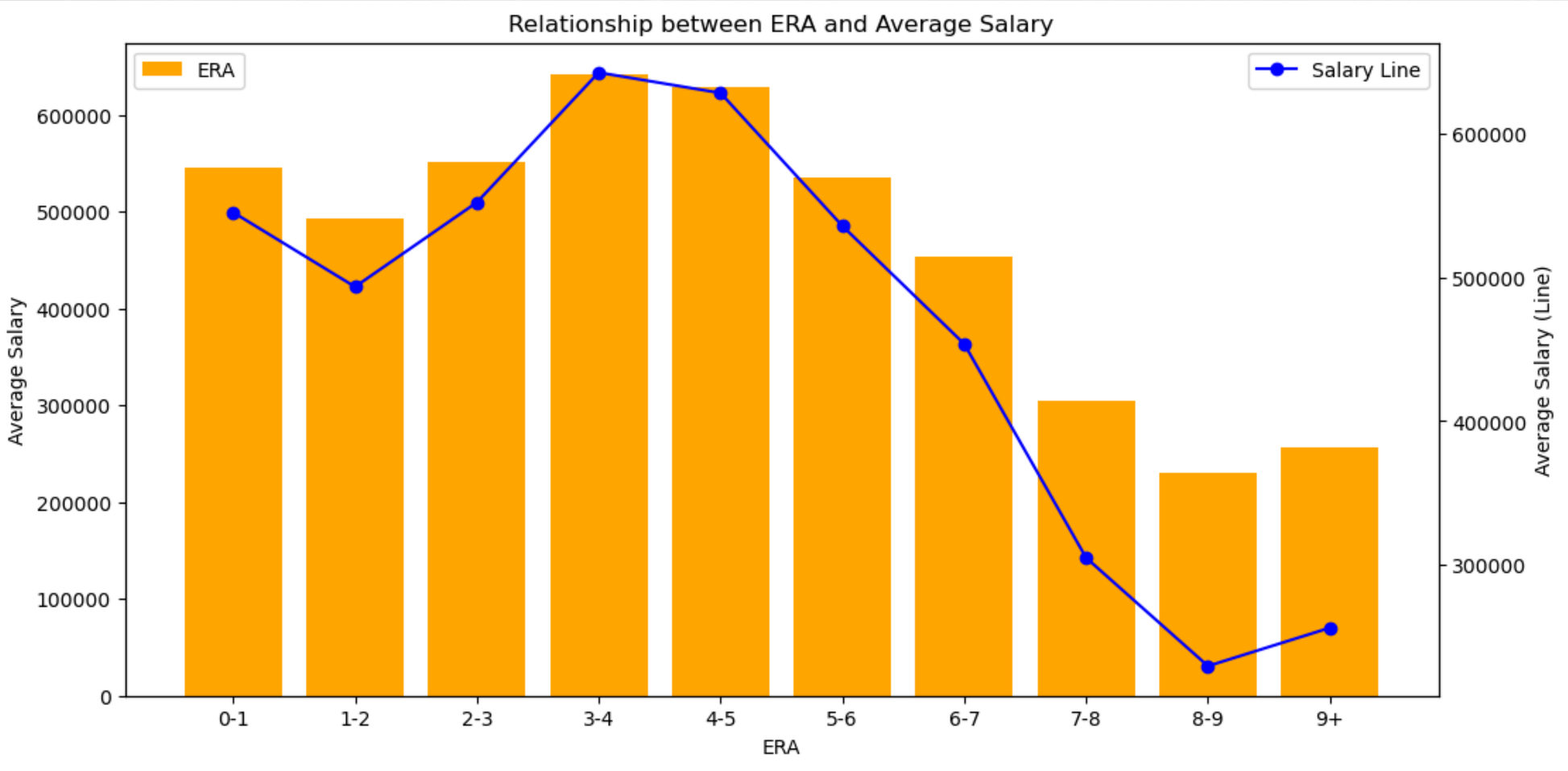
- Early Career Growth
- Peak Earning Age
- Decline After Prime

Salary & Home Run Relationship

- Higher Home Runs, Higher Salary
- Significant Salary Jumps at Higher Home Runs
- Performance Incentive



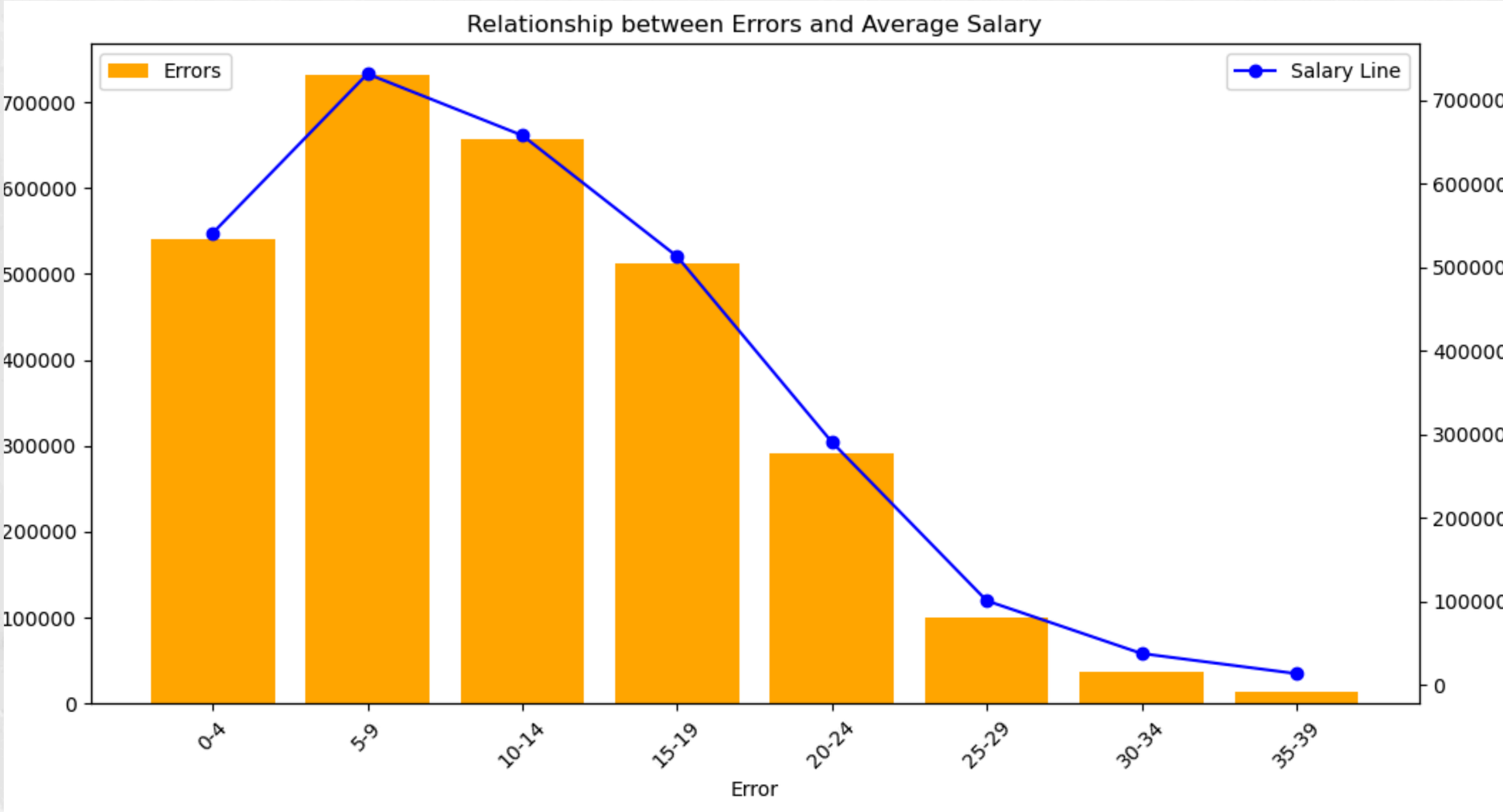
Salary & Earned Run Average Relationship



- Lower ERA, Higher Salary
- Mid-Range Peaks
- Decline at Higher ERA

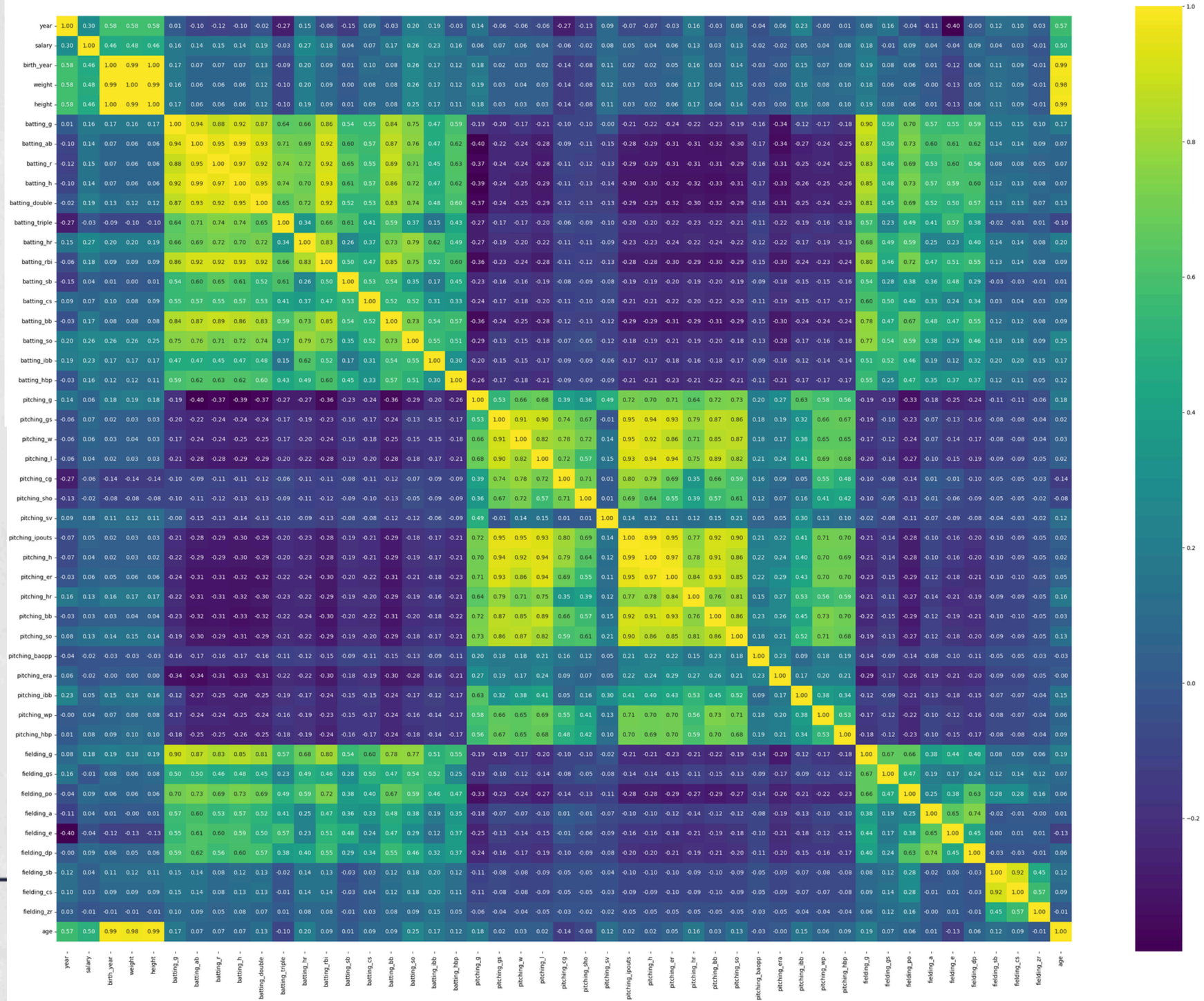
Salary & Errors Relationship

- Negative Correlation Between Errors and Salary
- Significant Salary Drop with Increased Errors
- Peak Salary for Moderate Error Ranges



MLB

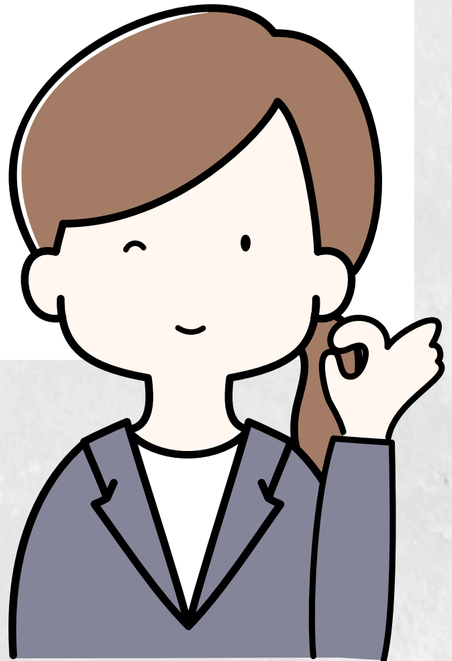
Correlation Matrix



Correlation Matrix

```
High Correlation with Target Variable:
salary    1.0
Name: salary, dtype: float64

Medium Correlation with Target Variable:
year           0.301468
birth_year     0.460520
weight         0.477073
height         0.461674
age            0.500996
Name: salary, dtype: float64
```



```
Low Correlation with Target Variable:
batting_g          0.162585
batting_ab         0.143519
batting_r          0.151575
batting_h          0.144799
batting_double     0.187518
batting_triple     0.025568
batting_hr         0.268987
batting_rbi        0.183750
batting_sb         0.038129
batting_cs         0.065881
batting_bb         0.171244
batting_so         0.262054
batting_ibb        0.233994
batting_hbp        0.163760
pitching_g         0.058960
pitching_gs        0.074789
pitching_w         0.059086
pitching_l         0.040721
pitching_cg        0.058503
pitching_sho       0.022662
pitching_sv        0.081647
pitching_ipouts    0.047213
pitching_h         0.043255
pitching_er        0.056441
pitching_hr        0.125930
pitching_bb        0.028902
pitching_so        0.128683
pitching_baopp     0.017942
pitching_era       0.024174
pitching_ibb       0.047113
pitching_wp        0.044264
pitching_hbp       0.083586
fielding_g         0.175488
fielding_gs        0.014045
fielding_po        0.094696
fielding_a         0.040545
fielding_e         0.043460
fielding_dp        0.093337
fielding_sb        0.040331
fielding_cs        0.026361
fielding_zr        0.013886
Name: salary, dtype: float64
```

Target Variable “Salary”

03 Machine Learning Models

03-Machine Learning Models

Naive Baseline Model

```
The Naive Baseline Model's MSE on train data is 3775465.11.
The Naive Baseline Model's MSE on test data is 3714239.63.
```

Linear Regression Model

```
The LinReg Model's MSE on train data is 1350056.5600.
The LinReg Model's MSE on test data is 1498301.3953.
```

MSE Change

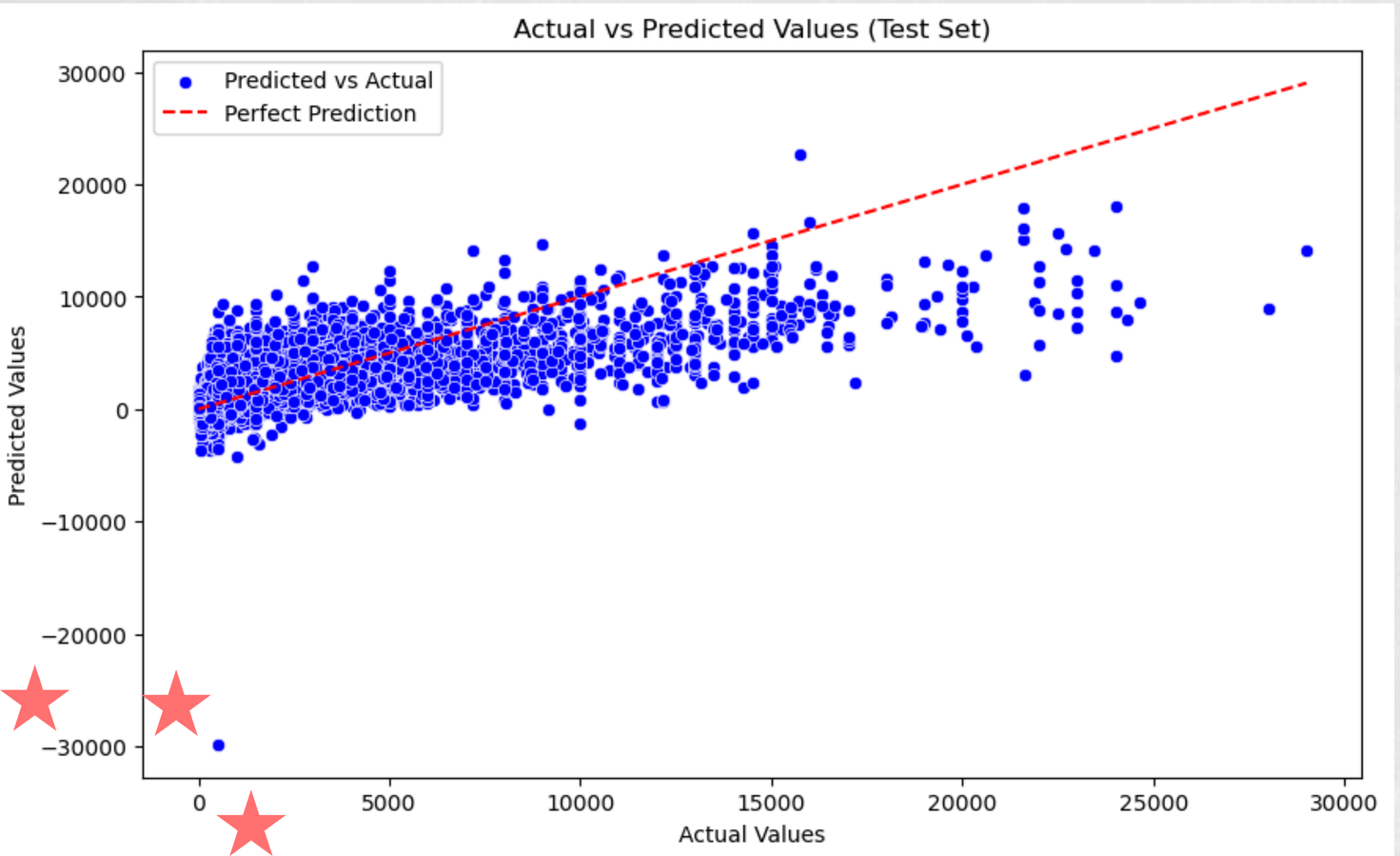
- Train Data: MSE decreased by approximately 64.23%
- Test Data: MSE decreased by approximately 59.64%

What does this mean?

Better performance and prediction score.

03-Model Performance

```
The LinReg Model's R2 on train data is 0.6424.
The LinReg Model's R2 on test data is 0.5966.
```



The Linear Regression Model has an R^2 value of approximately 0.6424 on train data and 0.5966 on test data, showing that it explains around 60% of the variability in salaries.

Blue Dots (Predicted vs. Actual)
Each blue dot represents a player, showing the relationship between their actual salary and the salary predicted by the model.

Red Dashed Line (Perfect Prediction)
The red dashed line represents a perfect prediction scenario where the predicted salaries would match the actual salaries exactly.

04
Conclusion

04-Conclusion

Data Insights Gathered

- Players with higher counts in key performance metrics such as Home Runs, Runs Batted In (RBIs), and Wins generally command higher salaries. This trend highlights the value teams place on strong offensive and pitching performance.

- Player salaries tend to peak in the age range of 30-39. Younger players (under 25) and older players (over 40) generally earn less, possibly due to perceived potential and physical decline, respectively.

- Players with better fielding metrics, such as fewer Errors and more Assists and Putouts, tend to earn higher salaries. This suggests that defensive skills are an important factor in salary determination.

04-Conclusion

Recommendations

- Consider structuring contracts that heavily incentivize performance metrics such as Home Runs, low errors, and low ERA. This approach not only rewards top-performing players but also aligns their incentives with team success.

- Recruitment and scouting efforts should prioritize players in the 30-39 age range, where performance and experience peak. However, teams should also strategically invest in younger players to ensure a pipeline of talent for the future.

- Given the importance of fielding performance metrics like Assists and Putouts, teams should enhance defensive training programs. This can help improve players' defensive skills, making them more valuable and potentially increasing their salaries.

Thanks!

CREDITS: This presentation template was created by **Slidesgo**, and includes icons by **Flaticon**, and infographics & images by **Freepik**