BASEBALL ANALYSIS

"The MVP of Baseball"
Data Analytics in Baseball Strategy

Project Sprint 2 by group #6

AJAY, MAYBANKS, JORDAN

https://youtu.be/1-wFL3k3knM

Git repository

AGENDA

First Base

- *Introduction & Business Objectives*

Second Base

- *Model Design & Development*

Third Base

- *Data's Effect On Baseball Eras*

Home Plate

- *Model Forecast & Validation*
- *Conclusions*

First Base
Introduction & Business Objectives

Introduction

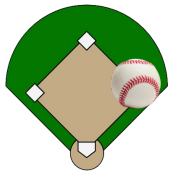
In the world of baseball, **success** is not only determined by the **skill of individual players**, but also by a range of **key variables** that can **impact** the outcome of a game.

- **Analyze data** that spans the **history** of major league baseball teams.

- Use **Statistical Analysis** to provide an insight into the **factors** that will bring home a **win**.

- Use **Modelling** & Examine **Variable Contribution**
  - Apply modelling
  - Evaluate
  - Make predictions

First Base



Business Objectives

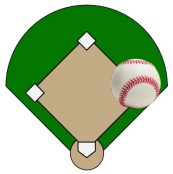
Determine Team & Franchise success using:

- Offensive statistics
- Use **Metrics from All Eras**
- Defensive Statistics
- **Combination** of both **Offense & Defense**

Analytical Objectives

- Conduct a **correlation analysis** to identify the **best variables**
- Test and train our **linear regression models** for all time periods
- Use metrics to evaluate our **Model accuracy**
- **Forecast** the number of **games won** by the **NY** Yankees and **Toronto** Blue Jays.

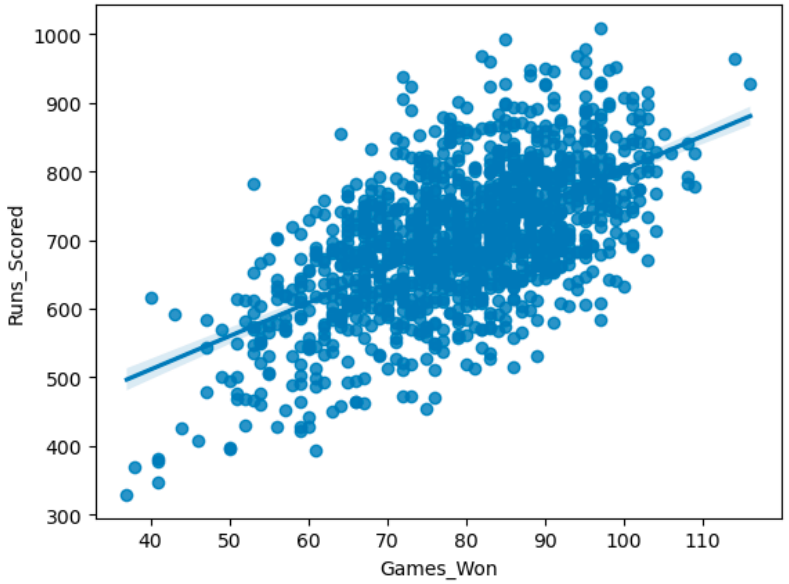
First Base

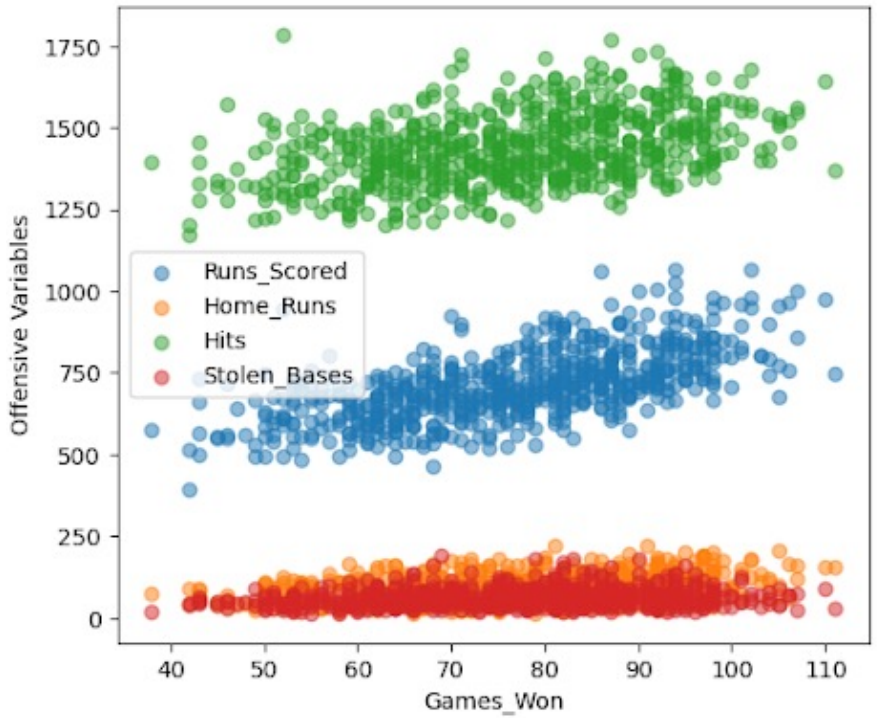


Second Base
Model Design & Development

MLR Models

Going from Linear Regression to Multiple Linear Regression





'Games_Won',

'Games_Lost',

'Runs_Against',

'Runs_Scored',

'Home_Runs',

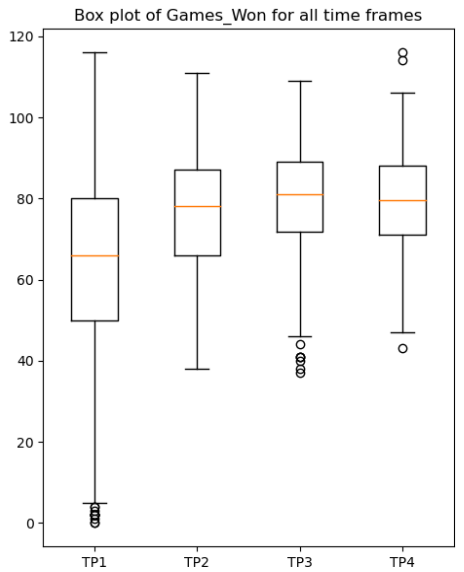
'Home_Run_Allowed',

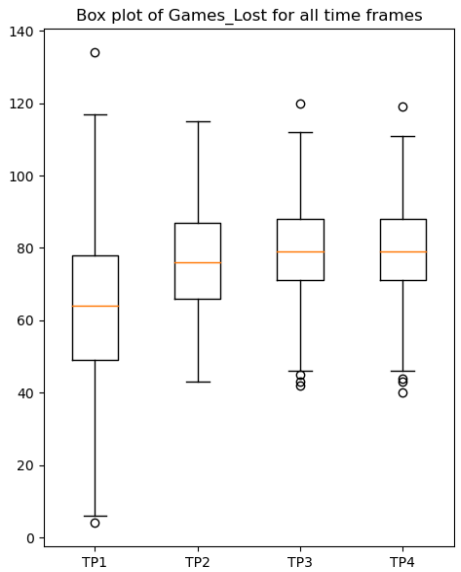
'Hits', 'Strikeouts_Allowed'

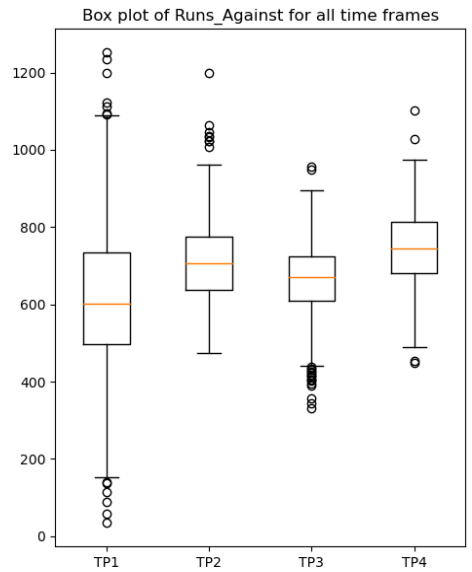
Second Base

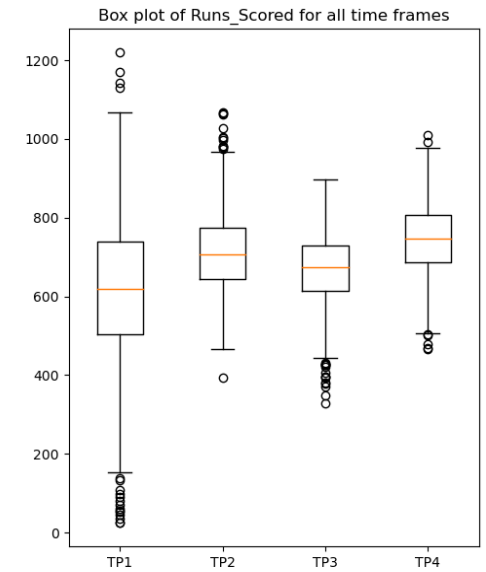
Old Variables

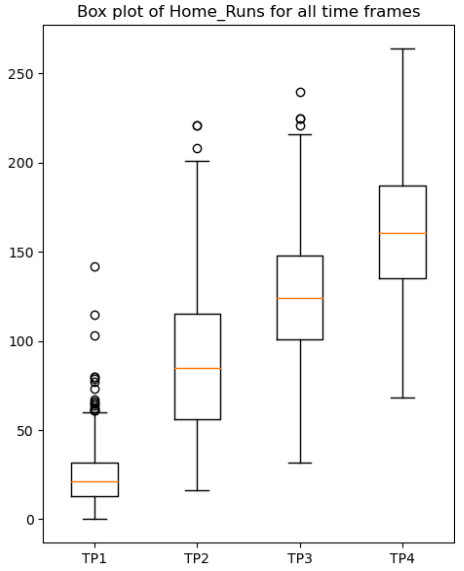
Boxplots for 8 Variables (1871-2010)

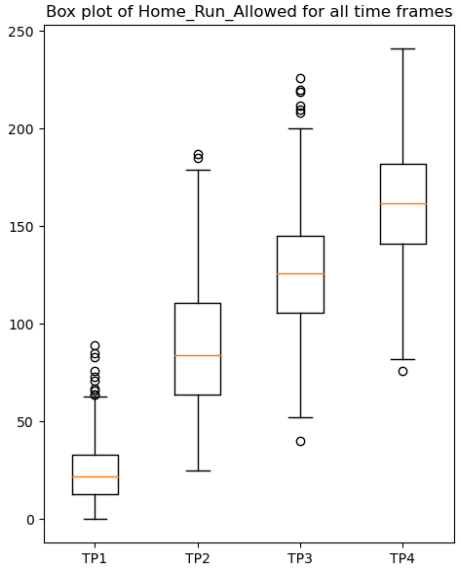


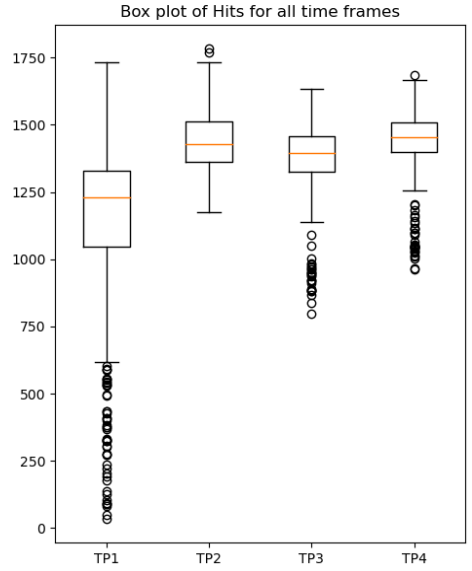


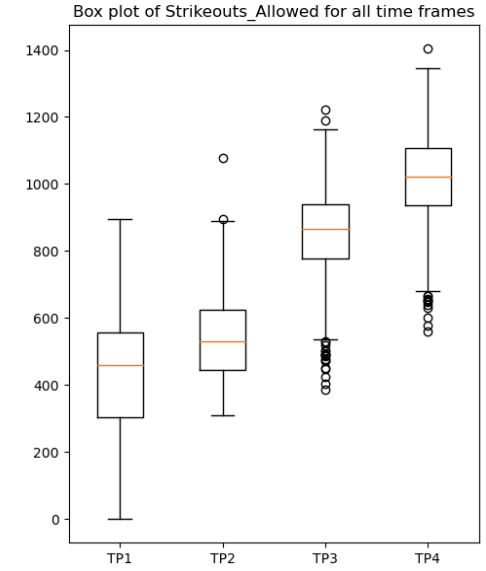












'Games_Won',

'Games_Lost',

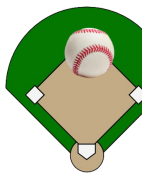
'Runs_Against',

'Runs_Scored',

'Home_Runs',

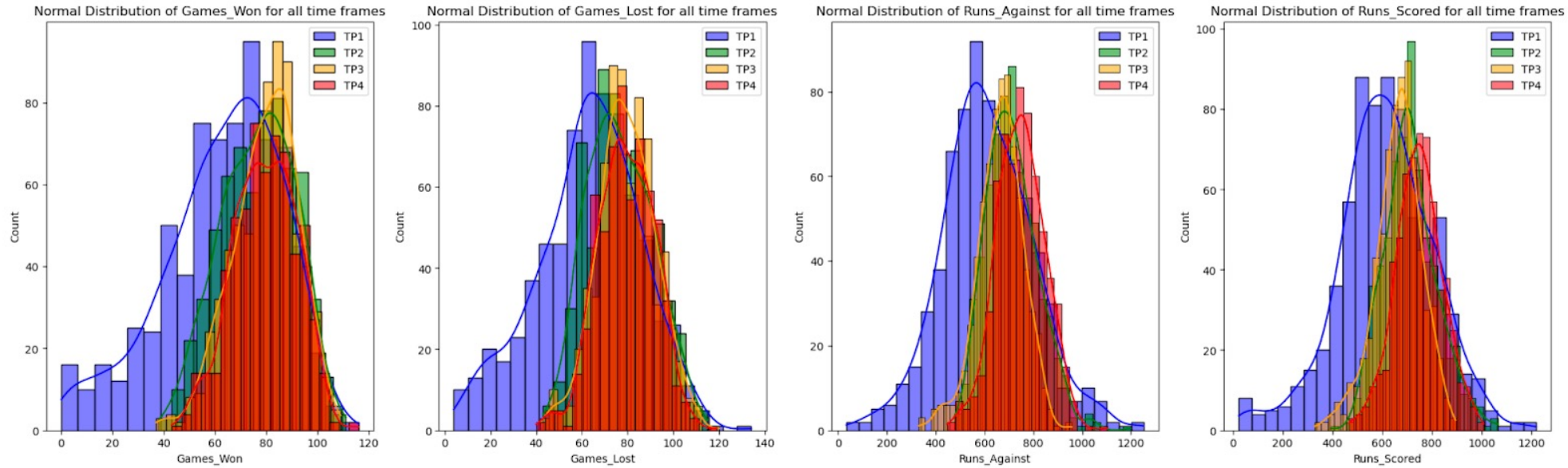
'Home_Run_Allowed',

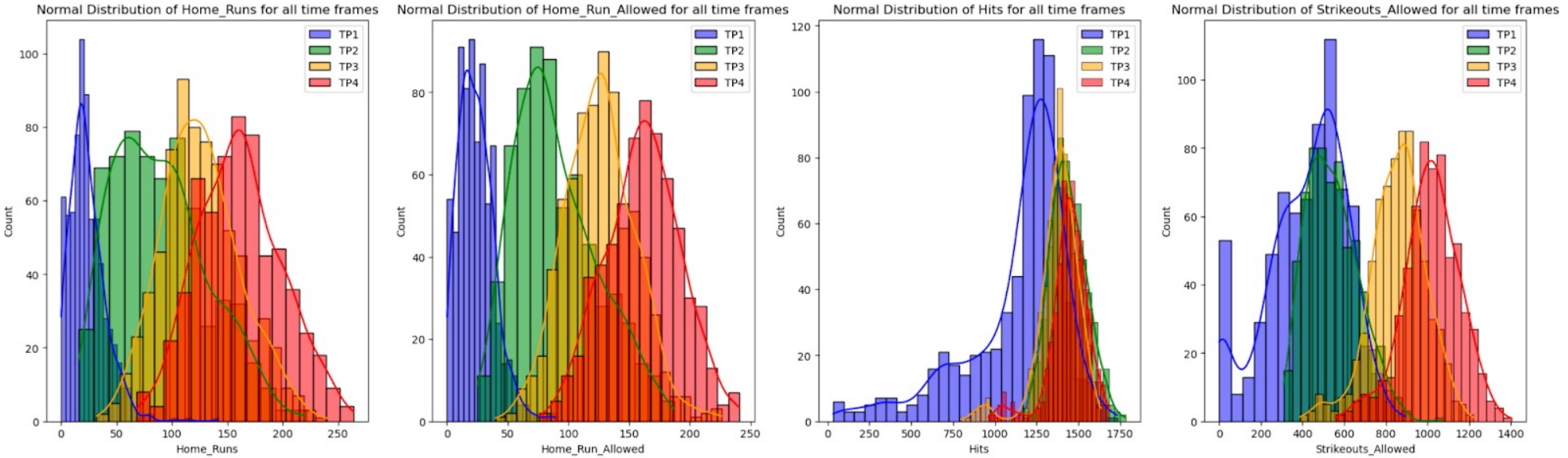
'Hits', 'Strikeouts_Allowed'



Old Variables

Distribution of Selected 8 Variables (1871-2010)





'Games_Won',

'Games_Lost',

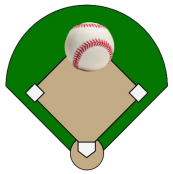
'Runs_Against',

'Runs_Scored',

'Home_Runs',

'Home_Run_Allowed',

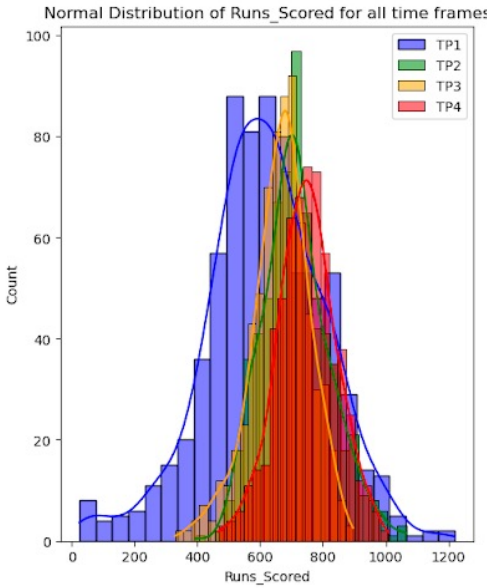
'Hits', 'Strikeouts_Allowed'

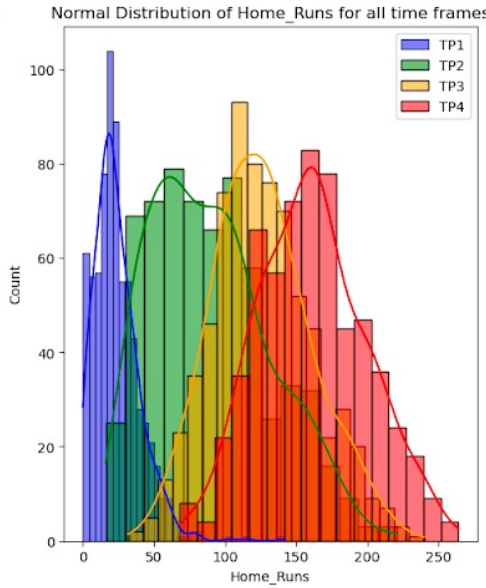


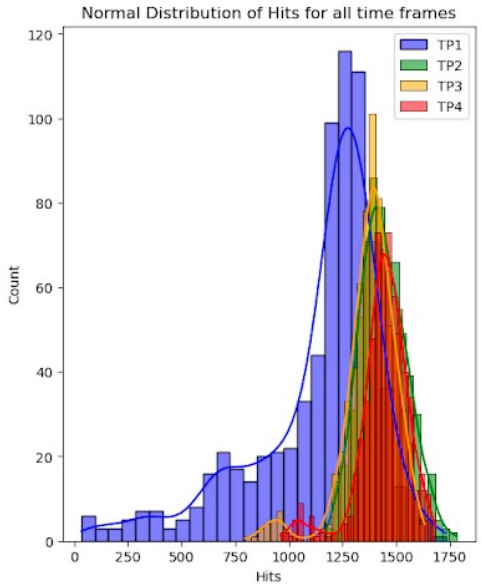
New Variables

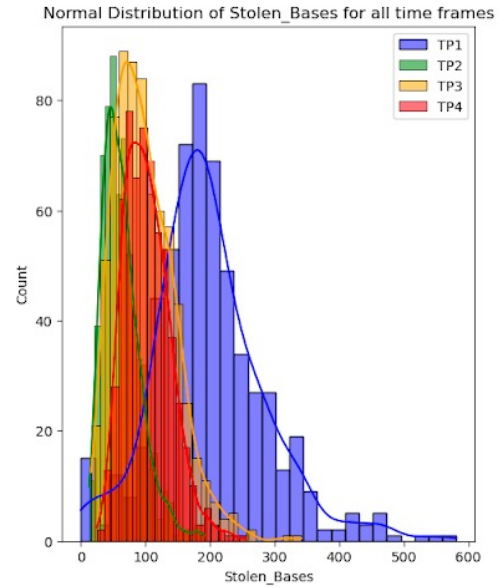
Distribution of New 8 Variables Across 4 time periods

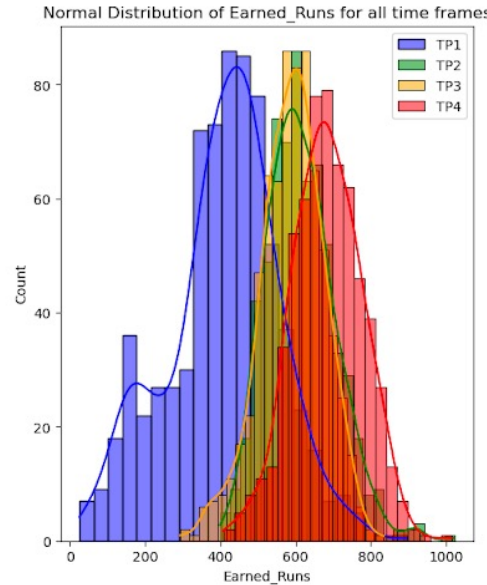
Second Base

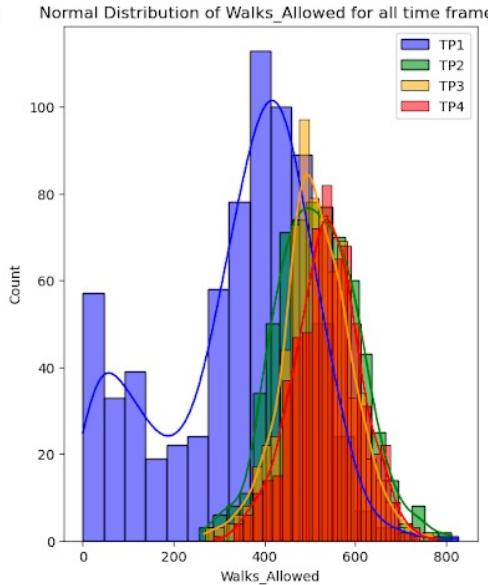


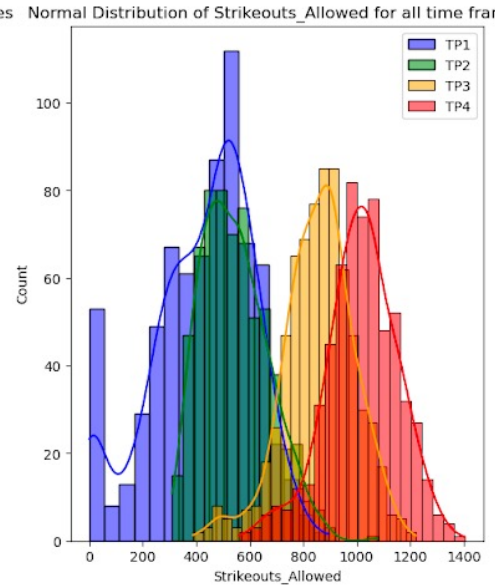


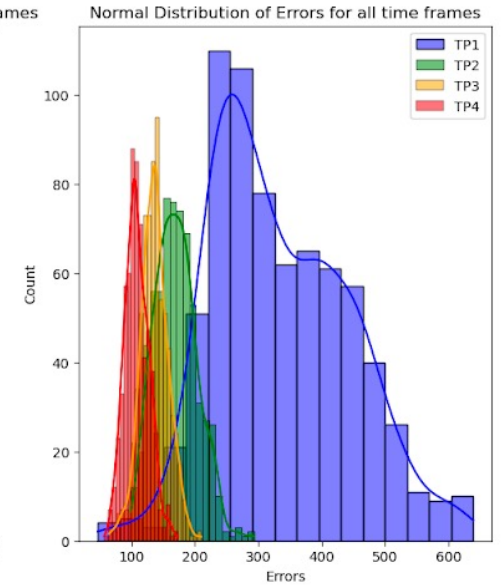












- 'Runs_Scored', 'Home_Runs', 'Hits', 'Stolen_Bases',
- 'Earned_Runs', ' Walks_Allowed', 'Strikeouts_Allowed', 'Errors'

Third Base

Data's Effect On Baseball Eras

Dead Ball Era (before 1920)

- During this time, the baseball used in games was less lively and harder to hit far, resulting in fewer home runs and more emphasis on small-ball tactics like bunting, stealing bases, and hitting for contact.

- The era is often associated with a style of play that was more defensive-oriented, with a premium placed on strong pitching, fielding, and base running. Games were often low-scoring affairs, and the style of play was more focused on strategy and finesse than power and speed.

Third Base

- It wasn't until the emergence of players like Ty Cobb and Babe Ruth in the 1910s that attendance really started to skyrocket. Before that, attendance varied depending on a number of factors such as team success, regional popularity, and the quality of the stadium.

- Some teams during the Dead Ball Era were very successful, while others struggled. For example, the New York Giants won four National League pennants and one World Series during the era, while the St. Louis Browns finished in last place in the American League nine times. Teams that were successful on the field tended to draw larger crowds, but that wasn't always the case.

Lively Ball Era (1910 - 1960)

- The 1920s saw record-breaking attendance numbers, with over 10 million fans attending games in 1920 and over 11 million in 1929.
- Offensive explosion resulted from the introduction of the "lively" ball and rule changes, making home runs and high-scoring games the norm.
- Success and attendance were not evenly distributed across all franchises. Smaller market teams and those in less populous regions struggled to draw large crowds and compete with the powerhouse teams.
- In terms of franchise success, the Yankees dominated the era, winning six World Series championships between 1923 and 1941.

Third Base

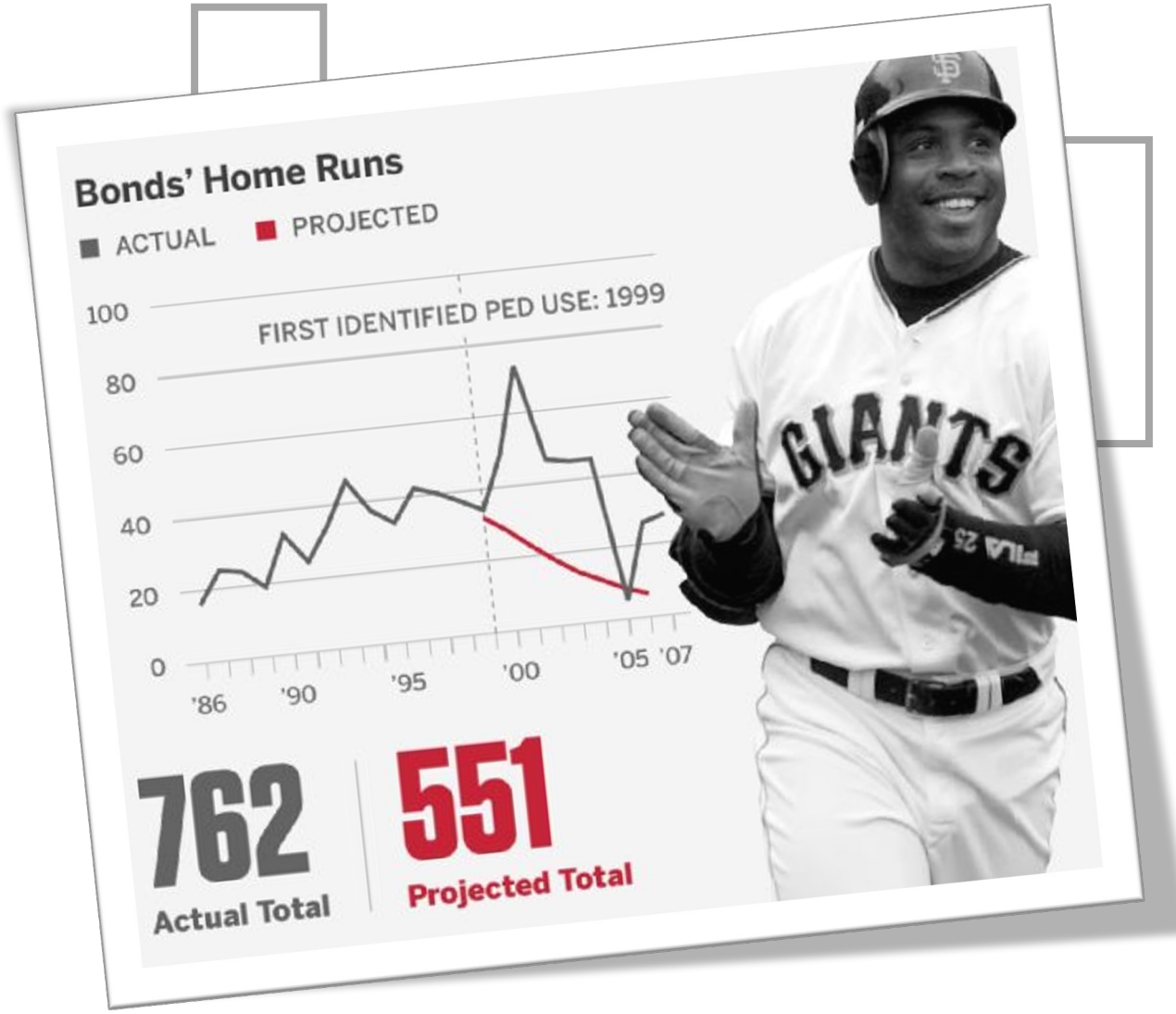
Expansion Era (1960-1990)

- The league underwent significant expansion by adding new teams. The addition of new teams during the Expansion Era had a profound impact on the league as it increased the number of games played each season, created new rivalries, and expanded the league's reach into new regions of the country. It also provided an opportunity for new players to enter the league and for existing players to play for new teams.
- The Expansion Era also coincided with changes in the game, such as the introduction of artificial turf and the designated hitter rule in the AL. These changes had a significant impact on the way the game was played and have continued to shape the sport to this day.
- Overall, the addition of new teams during the Expansion Era created a mix of successful and struggling franchises, with attendance often reflecting the success of the teams on the field. Successful teams tended to draw larger crowds, while struggling teams often had trouble filling their stadiums.

Third Base

Steroid Era (1990-2004)

- During this era, performance-enhancing drugs (PEDs) were widely used by players. The use of PEDs was believed to have contributed to an increase in offensive output and home runs during this period.
- In terms of attendance, the steroid era saw a significant increase in popularity, with record-breaking attendance figures in the late 1990s and early 2000s. Fans were drawn to the exciting offensive performances and home runs, and many of baseball's most famous players, such as Mark McGwire, Sammy Sosa, and Barry Bonds, became household names.
- In terms of franchise success, the steroid era saw mixed results. While some teams that were able to field rosters with steroid-enhanced players saw success on the field, the use of PEDs also contributed to a culture of cheating and undermined the integrity of the game. The success of teams during this period has been somewhat tarnished by revelations of widespread PED use by players.
- Overall, the steroid era saw a significant increase in attendance and a shift towards an offensive-oriented style of play. However, the use of PEDs also contributed to a culture of cheating and undermined the integrity of the game, leading to a decline in popularity and public trust in baseball.



Third Base

OLS Regression Results

Third Base

Third Base

Time Period 1: 1871 - 1920

All Variables:

OLS Regression Results

Dep. Variable:	Games_Won	R-squared:	0.919			
Model:	OLS	Adj. R-squared:	0.918			
Method:	Least Squares	F-statistic:	1005.			
Date:	Tue, 04 Apr 2023	Prob (F-statistic):	0.00			
Time:	17:14:46	Log-Likelihood:	-2375.9			
No. Observations:	719	AIC:	4770.			
Df Residuals:	710	BIC:	4811.			
Df Model:	8					
Covariance Type:	nonrobust					

	coef	std err	t	P>|t|	[0.025	0.975]
const	5.7310	1.424	4.023	0.000	2.934	8.528
Runs_Scored	0.0483	0.004	13.052	0.000	0.041	0.056
Home_Runs	-0.0174	0.022	-0.805	0.421	-0.060	0.025
Hits	0.0535	0.003	19.598	0.000	0.048	0.059
Stolen_Bases	0.0155	0.004	4.380	0.000	0.009	0.022
Earned_Runs	-0.0866	0.004	-21.764	0.000	-0.094	-0.079
Walks_Allowed	-0.0013	0.004	-0.364	0.716	-0.009	0.006
Strikeouts_Allowed	0.0335	0.002	15.779	0.000	0.029	0.038
Errors	-0.0414	0.003	-12.900	0.000	-0.048	-0.035

Omnibus:	0.762	Durbin-Watson:	1.834
Prob(Omnibus):	0.683	Jarque-Bera (JB):	0.800
Skew:	0.078	Prob(JB):	0.670
Kurtosis:	2.951	Cond. No.	8.99e+03

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 8.99e+03. This might indicate that there are strong multicollinearity or other numerical problems.

Time Period 2: 1920 - 1960

Offensive Variables:

OLS Regression Results

Dep. Variable:	Games_Won	R-squared (uncentered):	0.980			
Model:	OLS	Adj. R-squared (uncentered):	0.980			
Method:	Least Squares	F-statistic:	7941.			
Date:	Tue, 04 Apr 2023	Prob (F-statistic):	0.00			
Time:	17:14:46	Log-Likelihood:	-2507.5			
No. Observations:	656	AIC:	5023.			
Df Residuals:	652	BIC:	5041.			
Df Model:	4					
Covariance Type:	nonrobust					

	coef	std err	t	P>|t|	[0.025	0.975]
Runs_Scored	0.0708	0.007	9.677	0.000	0.056	0.085
Home_Runs	0.0682	0.013	5.209	0.000	0.043	0.094
Hits	0.0116	0.003	3.612	0.000	0.005	0.018
Stolen_Bases	0.0505	0.017	2.912	0.004	0.016	0.085

Omnibus:	5.134	Durbin-Watson:	1.933
Prob(Omnibus):	0.077	Jarque-Bera (JB):	4.956
Skew:	-0.197	Prob(JB):	0.0839
Kurtosis:	3.164	Cond. No.	71.5

Notes:
[1] R^2 is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Time Period 3: 1960 - 1990

Offensive Variables:

OLS Regression Results

Dep. Variable:	Games_Won	R-squared (uncentered):	0.986			
Model:	OLS	Adj. R-squared (uncentered):	0.986			
Method:	Least Squares	F-statistic:	1.321e+04			
Date:	Tue, 04 Apr 2023	Prob (F-statistic):	0.00			
Time:	17:14:47	Log-Likelihood:	-2670.5			
No. Observations:	730	AIC:	5349.			
Df Residuals:	726	BIC:	5367.			
Df Model:	4					
Covariance Type:	nonrobust					

	coef	std err	t	P>|t|	[0.025	0.975]
Runs_Scored	0.0818	0.010	8.534	0.000	0.063	0.101
Home_Runs	-0.0065	0.017	-0.390	0.697	-0.039	0.026
Hits	0.0172	0.004	4.760	0.000	0.010	0.024
Stolen_Bases	0.0211	0.008	2.519	0.012	0.005	0.038

Omnibus:	1.425	Durbin-Watson:	1.816
Prob(Omnibus):	0.490	Jarque-Bera (JB):	1.389
Skew:	-0.025	Prob(JB):	0.499
Kurtosis:	2.792	Cond. No.	83.3

Notes:
[1] R^2 is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Time Period 4: 1990 - 2010

Defensive Variables:

OLS Regression Results

Dep. Variable:	Games_Won	R-squared (uncentered):	0.980			
Model:	OLS	Adj. R-squared (uncentered):	0.979			
Method:	Least Squares	F-statistic:	7224.			
Date:	Tue, 04 Apr 2023	Prob (F-statistic):	0.00			
Time:	17:14:48	Log-Likelihood:	-2347.8			
No. Observations:	608	AIC:	4704.			
Df Residuals:	604	BIC:	4721.			
Df Model:	4					
Covariance Type:	nonrobust					

	coef	std err	t	P>|t|	[0.025	0.975]
Earned_Runs	0.0002	0.006	0.037	0.971	-0.012	0.012
Walks_Allowed	-0.0070	0.008	-0.830	0.407	-0.023	0.010
Strikeouts_Allowed	0.0736	0.003	26.110	0.000	0.068	0.079
Errors	0.0704	0.024	2.930	0.004	0.023	0.118

Omnibus:	1.764	Durbin-Watson:	1.899
Prob(Omnibus):	0.414	Jarque-Bera (JB):	1.595
Skew:	0.116	Prob(JB):	0.451
Kurtosis:	3.097	Cond. No.	69.6

OLS Regression Results

Third Base

Time Period 1: 1871 - 1920

All Variables:

OLS Regression Results

Dep. Variable:	Games_Won	R-squared:	0.919
Model:	OLS	Adj. R-squared:	0.918
Method:	Least Squares	F-statistic:	1005.
Date:	Tue, 04 Apr 2023	Prob (F-statistic):	0.00
Time:	17:14:46	Log-Likelihood:	-2375.9
No. Observations:	719	AIC:	4770.
Df Residuals:	710	BIC:	4811.
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	5.7310	1.424	4.023	0.000	2.934	8.528
Runs_Scored	0.0483	0.004	13.052	0.000	0.041	0.056
Home_Runs	-0.0174	0.022	-0.805	0.421	-0.060	0.025
Hits	0.0535	0.003	19.598	0.000	0.048	0.059
Stolen_Bases	0.0155	0.004	4.380	0.000	0.009	0.022
Earned_Runs	-0.0866	0.004	-21.764	0.000	-0.094	-0.079
Walks_Allowed	-0.0013	0.004	-0.364	0.716	-0.009	0.006
Strikeouts_Allowed	0.0335	0.002	15.779	0.000	0.029	0.038
Errors	-0.0414	0.003	-12.900	0.000	-0.048	-0.035

Omnibus:	0.762	Durbin-Watson:	1.834
Prob(Omnibus):	0.683	Jarque-Bera (JB):	0.800
Skew:	0.078	Prob(JB):	0.670
Kurtosis:	2.951	Cond. No.	8.99e+03

Notes:

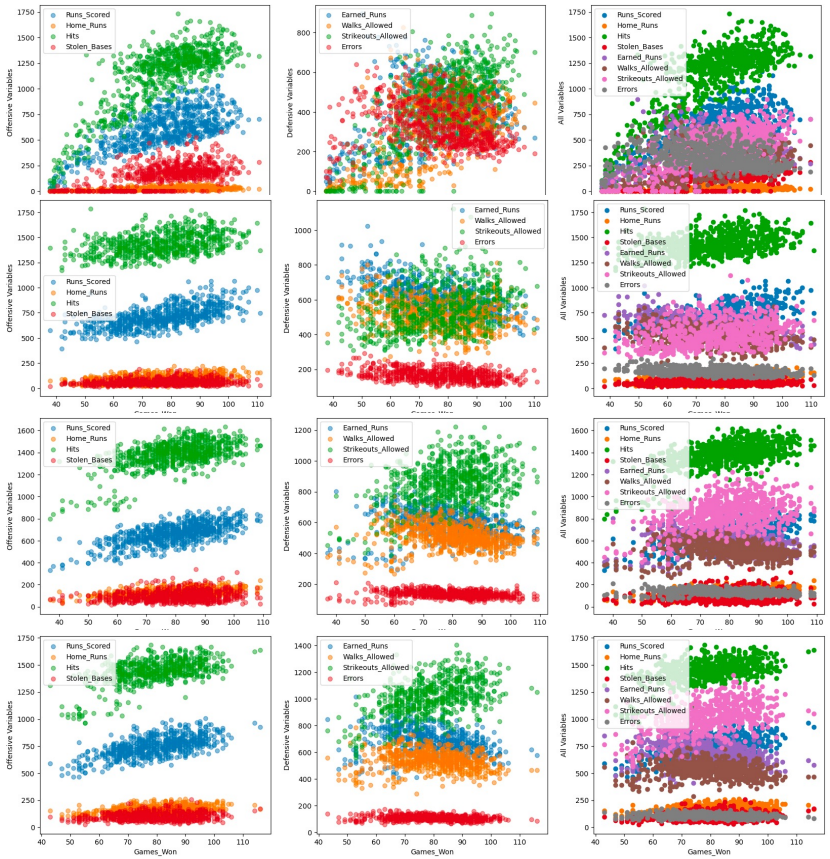
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

[2] The condition number is large, 8.99e+03. This might indicate that there are strong multicollinearity or other numerical problems.

How We Created The Models

12 Models

3 Groups (Offensive, Defensive, All) across 4 time periods.



Third Base

R-squared

		R-squared values		
	Time-frame1 r2	time-frame 2 r2	timeframe 3 r2	fime frame 4 r2
variables	0.919	0.908	0.851	0.845
def_variables	0.929	0.962	0.971	0.98
off_variables	0.959	0.98	0.986	0.986
		Adjusted R-squared values		
	Time-frame1 r2	time-frame 2 r2	timeframe 3 r2	fime frame 4 r2
variables	0.918	0.906	0.85	0.843
def_variables	0.928	0.962	0.971	0.979
off_variables	0.958	0.98	0.986	0.986

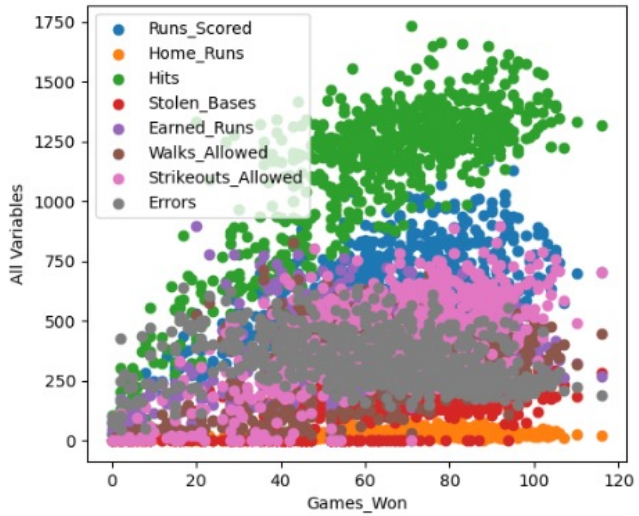
Third Base

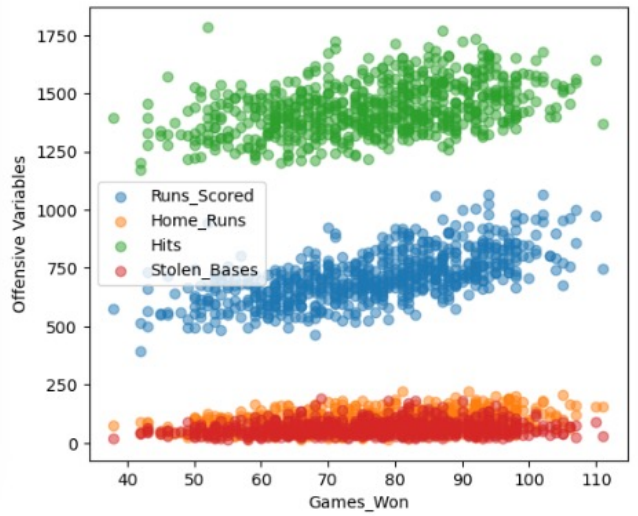
Home Plate

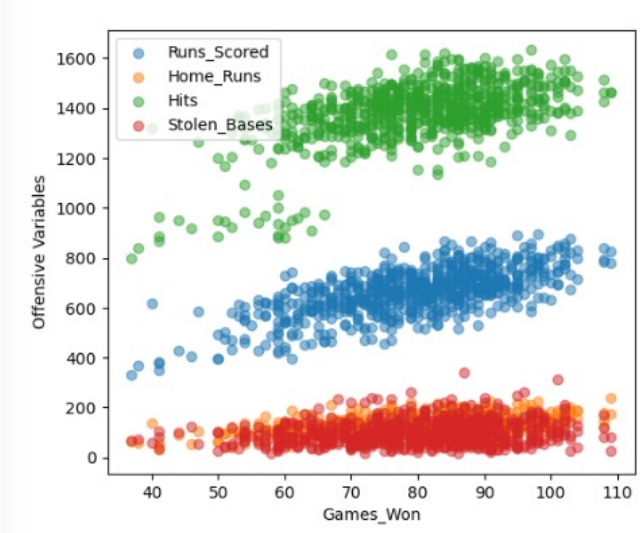
Model Validation, Forecast, & Conclusion

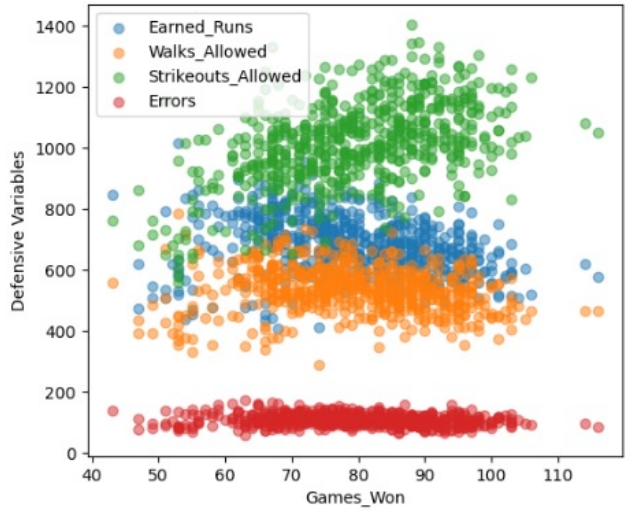
Model Validation

Home Plate



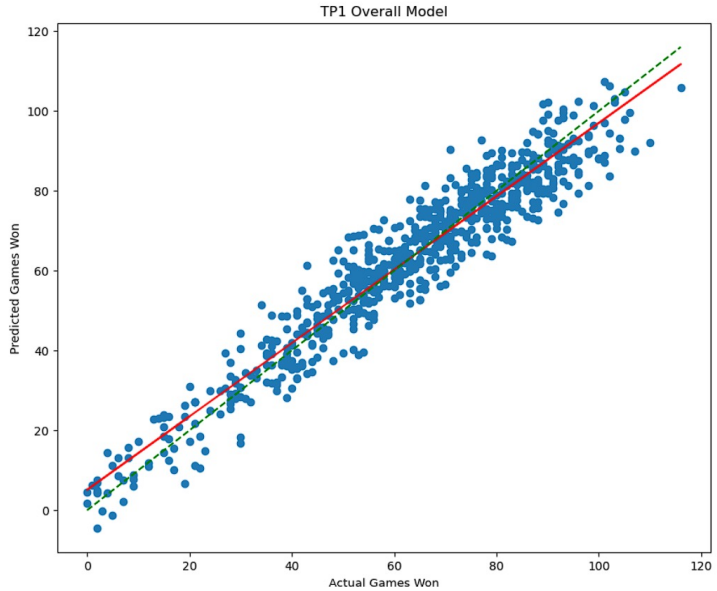


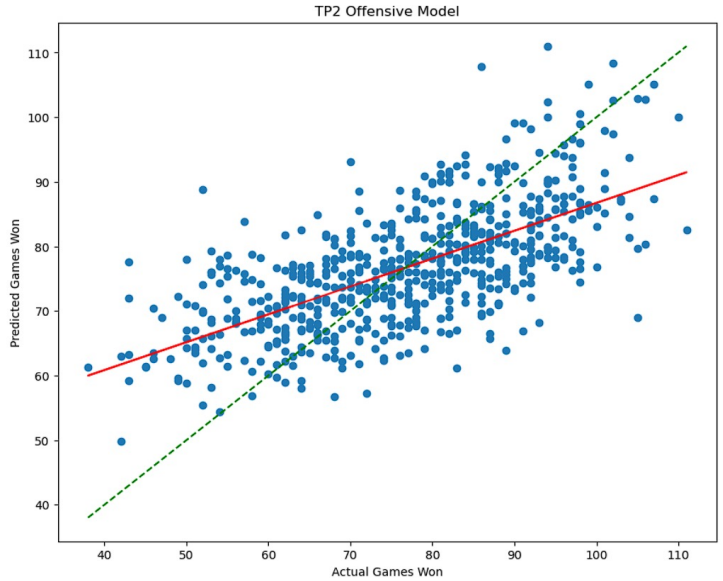


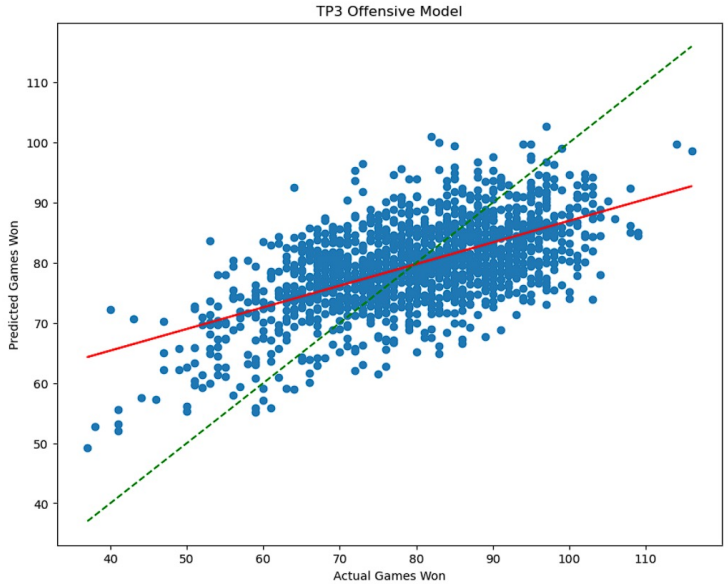


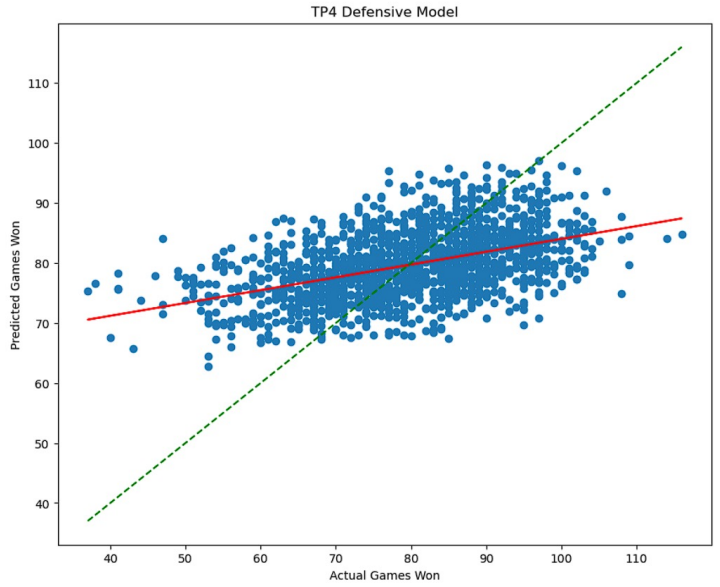
Model Validation

Home Plate



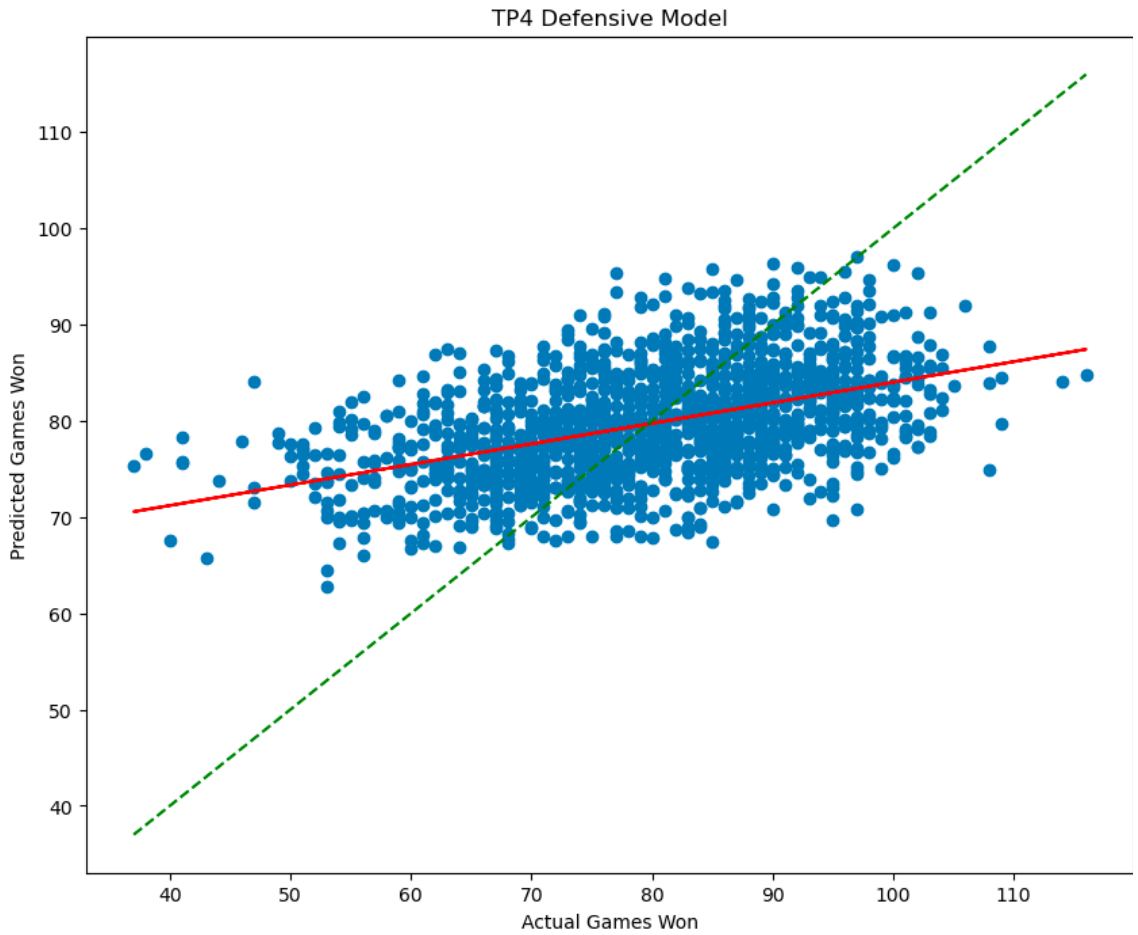






Model Validation

Using the 4th Regression model from 1990- 2010



Time Period 4: 1990 - 2010

Defensive Variables:

OLS Regression Results

Dep. Variable:	Games_Won	R-squared (uncentered):	0.980
Model:	OLS	Adj. R-squared (uncentered):	0.979
Method:	Least Squares	F-statistic:	7224.
Date:	Tue, 04 Apr 2023	Prob (F-statistic):	0.00
Time:	17:14:48	Log-Likelihood:	-2347.8
No. Observations:	608	AIC:	4704.
Df Residuals:	604	BIC:	4721.
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Earned_Runs	0.0002	0.006	0.037	0.971	-0.012	0.012
Walks_Allowed	-0.0070	0.008	-0.830	0.407	-0.023	0.010
Strikeouts_Allowed	0.0736	0.003	26.110	0.000	0.068	0.079
Errors	0.0704	0.024	2.930	0.004	0.023	0.118

Omnibus:	1.764	Durbin-Watson:	1.899
Prob(Omnibus):	0.414	Jarque-Bera (JB):	1.595
Skew:	0.116	Prob(JB):	0.451
Kurtosis:	3.097	Cond. No.	69.6

Home Plate

Home Plate

```
#using model 4 for yankees,2012 #actual_games_won = 95
import numpy as np

# Input values
input_stats = np.array([617, 431, 1318, 75]).reshape(1,-1)

# Predict games won
predicted_games_won = model4.predict(input_stats)

print("Predicted games won: ", predicted_games_won)
```

Predicted games won: [91.01696297]

```
#using model 4 for yankees,2015 #actual_games_won = 89
import numpy as np

# Input values
input_stats = np.array([652, 474, 1370, 93]).reshape(1,-1)

# Predict games won
predicted_games_won = model4.predict(input_stats)

print("Predicted games won: ", predicted_games_won)
```

Predicted games won: [90.77344009]

```
#using model 4 for toronto,2012 #actual_games_won = 73
import numpy as np

# Input values
input_stats = np.array([745, 574, 1142, 101]).reshape(1,-1)

# Predict games won
predicted_games_won = model4.predict(input_stats)

print("Predicted games won: ", predicted_games_won)
```

Predicted games won: [80.00176665]

```
#using model 4 for toronto,2015 #actual_games_won = 93
import numpy as np

# Input values
input_stats = np.array([609, 397, 1117, 88]).reshape(1,-1)

# Predict games won
predicted_games_won = model4.predict(input_stats)

print("Predicted games won: ", predicted_games_won)
```

Predicted games won: [85.98151026]

Prediction & Conclusions

Model	Team	Year	Actual Wins	Predicted Wins	Off By (Wins)
4	NYY	2012	95	91.01	4
4	TOR	2012	89	90.77	2
4	NYY	2015	73	80	7
4	TOR	2015	93	85.98	7

Home Plate

Root Mean Square Error

In [62]:
```
#using model 4 for yankees,2012 #actual_games_won = 95


import numpy as np
from sklearn.metrics import mean_squared_error

#using model 4 for yankees,2012 #actual_games_won = 95
mse1 = mean_squared_error([95], [91])
rmse1 = np.sqrt(mse1)

print("RMSE: ", rmse1)
```

RMSE: 4.0

In [66]:
```
#using model 4 for yankees,2015 #actual_games_won = 89
mse2 = mean_squared_error([89], [90.9])
rmse2 = np.sqrt(mse2)

print("RMSE: ", rmse2)
```

RMSE: 1.9000000000000057

In [68]:
```
#using model 4 for toronto,2012 #actual_games_won = 73
mse3 = mean_squared_error([73], [80])
rmse3 = np.sqrt(mse3)

print("RMSE: ", rmse3)
```

RMSE: 7.0

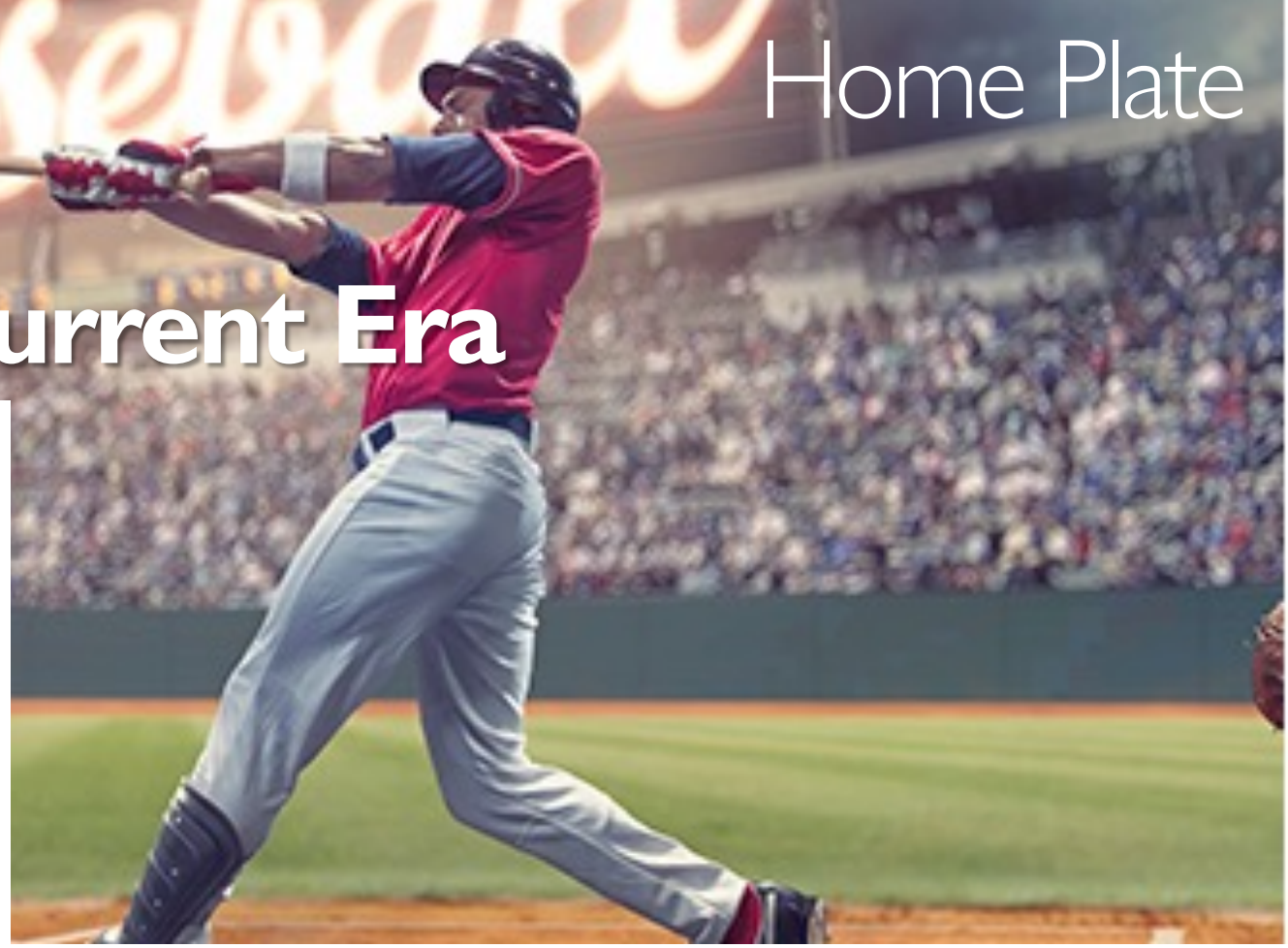
In [67]:
```
#using model 4 for toronto,2015 #actual_games_won = 93
mse4 = mean_squared_error([93], [86])
rmse4 = np.sqrt(mse4)

print("RMSE: ", rmse4)
```

RMSE: 7.0

Home Plate

Direction For The Current Era



+ Conclusion

- The accuracy of the model depends not only on the variables but also the time period.
- While some models are 98% accurate for some scenarios, it might not be so for all other scenarios.
- Adjustment of variables is required for improved accuracy depending upon the era and it's intricacies.

Home Plate


Thank you!