Analyzing Team Performance in Baseball

(1871 – 2015)

Matthew Martin

ISDA 150 – Sports Analytics

SJSU School of Information

May 11, 2024

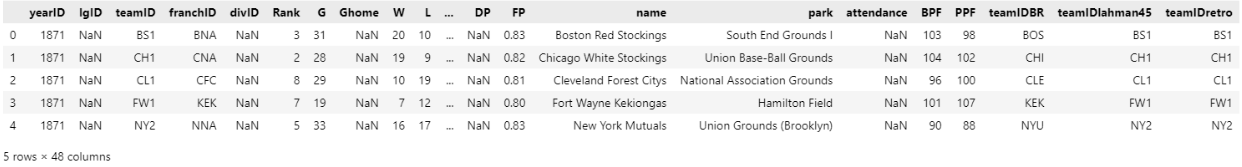
Overview

Baseball is one of the oldest sports played in the United States and is said to be “America’s Pastime.” In 1871, the ‘National Association of Professional Base Ball Players’ was founded, and introduced franchises that are still around today, including the Atlanta Braves and Chicago Cubs. Analyzing the player and team data will allow us to gain insights into the game and understand what factors contribute to success in baseball. The dataset I will be analyzing, ‘Baseball Databank’, includes yearly statistics and standings between 1871 and 2015. Throughout this report, I will be attempting to answer five major questions about baseball:

- What factors contribute to a team’s success?
- How does team performance vary across different leagues and divisions?
- Is there a correlation between attendance and team performance?
- How have park factors influenced team performance over time?
- Can we predict a team’s playoff success based on regular season performance?

Dataset Exploration

The main dataset that I used throughout this exploratory analysis was of yearly statistics and standings for each team from 1871 until 2015. There are 48 columns of information, including year, team name, final position in standings, and game data such as wins, losses, runs, ERA, etc. With this data, I will be able to find valuable insights about professional baseball over time.

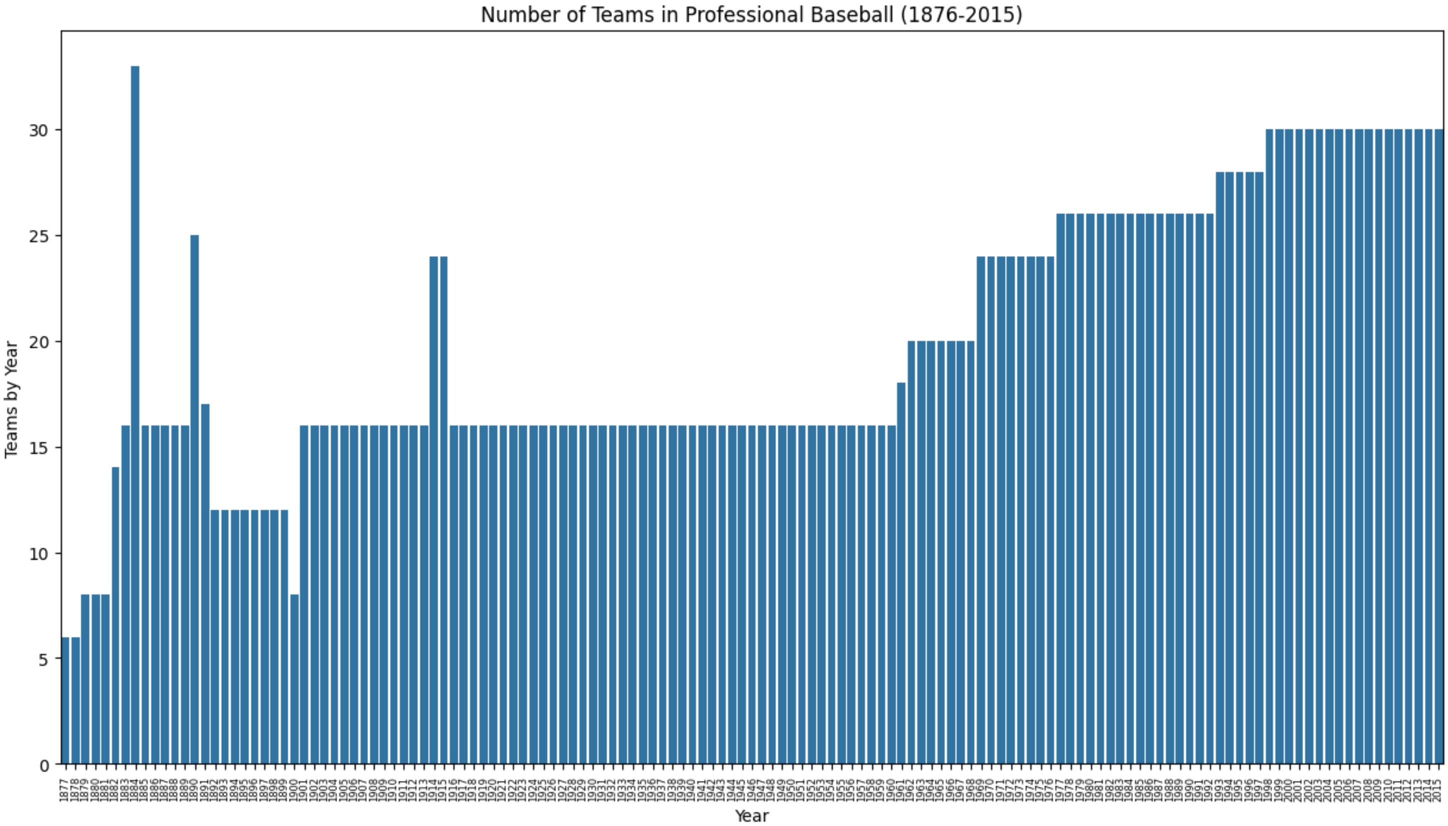


	yearID	lgID	teamID	franchID	divID	Rank	G	Ghome	W	L	...	DP	FP	name	park	attendance	BPF	PPF	teamIDBR	teamIDlahman45	teamIDretro
0	1871	NaN	BS1	BNA	NaN	3	31	NaN	20	10	...	NaN	0.83	Boston Red Stockings	South End Grounds I	NaN	103	98	BOS	BS1	BS1
1	1871	NaN	CH1	CNA	NaN	2	28	NaN	19	9	...	NaN	0.82	Chicago White Stockings	Union Base-Ball Grounds	NaN	104	102	CHI	CH1	CH1
2	1871	NaN	CL1	CFC	NaN	8	29	NaN	10	19	...	NaN	0.81	Cleveland Forest Citys	National Association Grounds	NaN	96	100	CLE	CL1	CL1
3	1871	NaN	FW1	KEK	NaN	7	19	NaN	7	12	...	NaN	0.80	Fort Wayne Kekiongas	Hamilton Field	NaN	101	107	KEK	FW1	FW1
4	1871	NaN	NY2	NNA	NaN	5	33	NaN	16	17	...	NaN	0.83	New York Mutuals	Union Grounds (Brooklyn)	NaN	90	88	NYU	NY2	NY2

5 rows × 48 columns

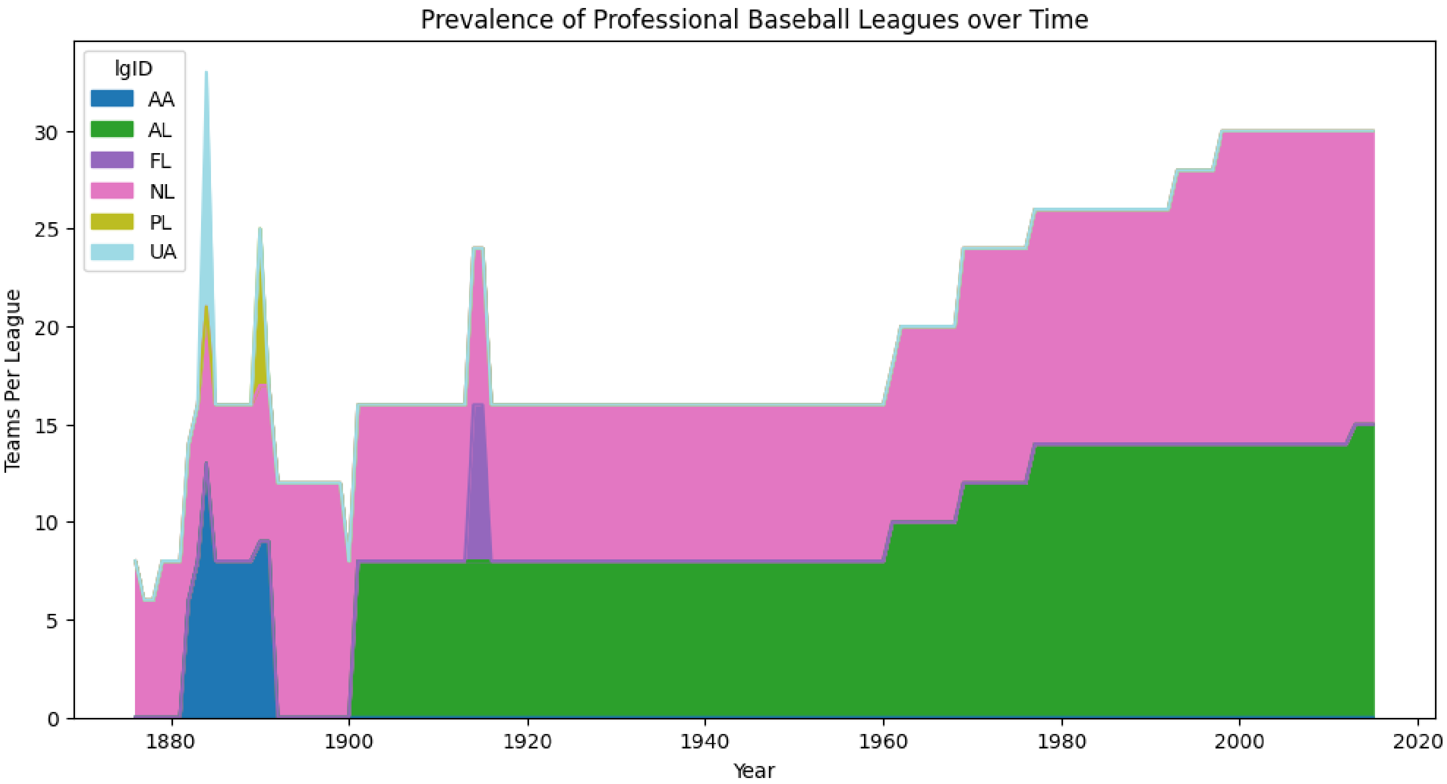
(Source: author)

The first insight that I explored with this dataset was looking at the number of total teams over time. Since this dataset dates back all the way to 1871, I was interested in seeing how it began and where it is now.



There is a big outlier in this data, which was in 1884. In 1884, the Major Leagues consisted of three different entities: National League, American Association, and the Union Association. These leagues consisted of 8, 13, and 12 teams respectively, totaling 33 teams. The Union Association only lasted one season, and the number of teams went back to normal in the years after. These days, there are 30 teams in the league, with 15 teams in each league (National League, American League).

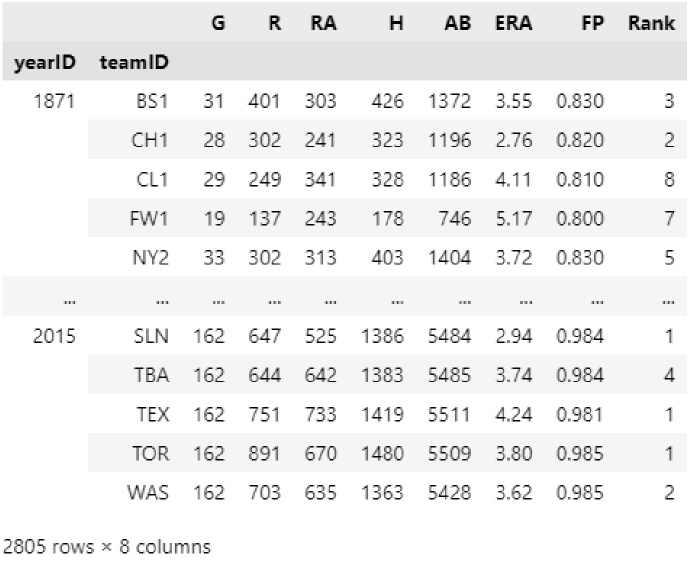
To see the prevalence of the different leagues over time, I plotted an area chart showing the number of teams in each league since 1871.



As the chart shows, there were a lot of leagues popping up and disbanding in the early years of professional baseball. The American Association was the most popular of these leagues, lasting from 1882 to 1891. The Union Association and the Players' League only lasted one season each, in 1884 and 1890 respectively. Finally, the Federal League was around for two seasons in 1914 and 1915 before the National League and American League solidified their spots as the two main leagues, which has held ever since.

Analysis and Insights

The first question that I set out to answer was which factors contribute most to a team's success. To do this, I gathered summary statistics for each year by team.

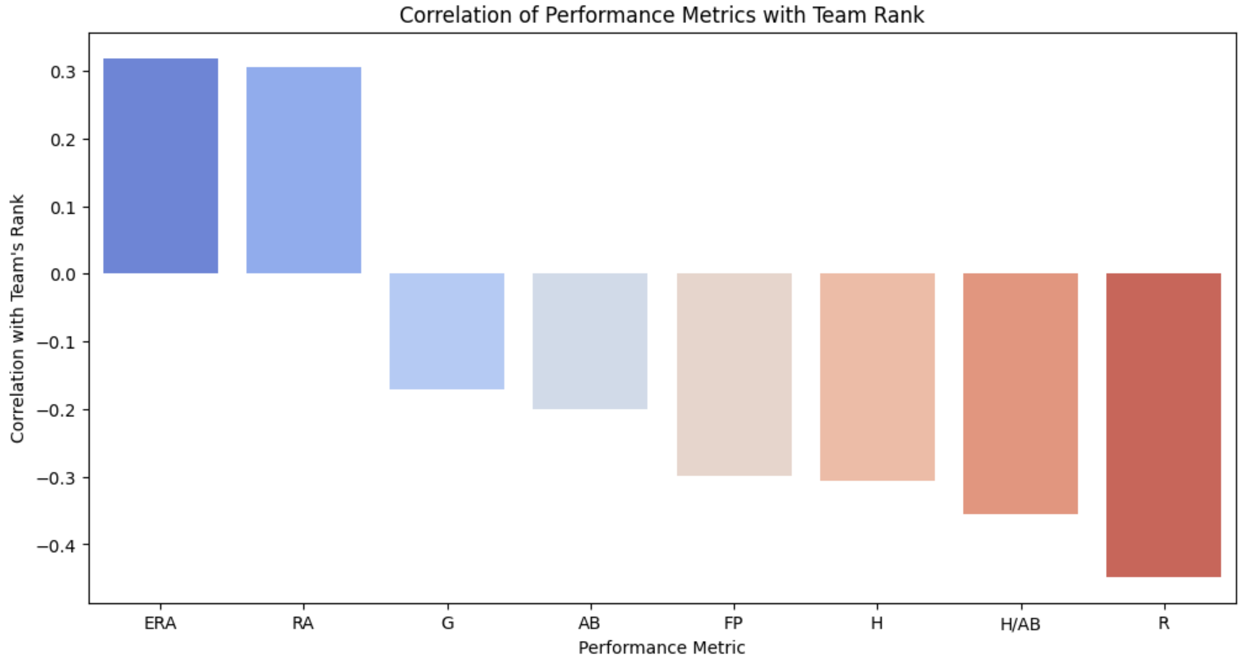


		G	R	RA	H	AB	ERA	FP	Rank
yearID	teamID								
1871	BS1	31	401	303	426	1372	3.55	0.830	3
	CH1	28	302	241	323	1196	2.76	0.820	2
	CL1	29	249	341	328	1186	4.11	0.810	8
	FW1	19	137	243	178	746	5.17	0.800	7
	NY2	33	302	313	403	1404	3.72	0.830	5
...	...	...	...	...	...	...	...	...	...
2015	SLN	162	647	525	1386	5484	2.94	0.984	1
	TBA	162	644	642	1383	5485	3.74	0.984	4
	TEX	162	751	733	1419	5511	4.24	0.981	1
	TOR	162	891	670	1480	5509	3.80	0.985	1
	WAS	162	703	635	1363	5428	3.62	0.985	2

2805 rows × 8 columns

(Source: author)

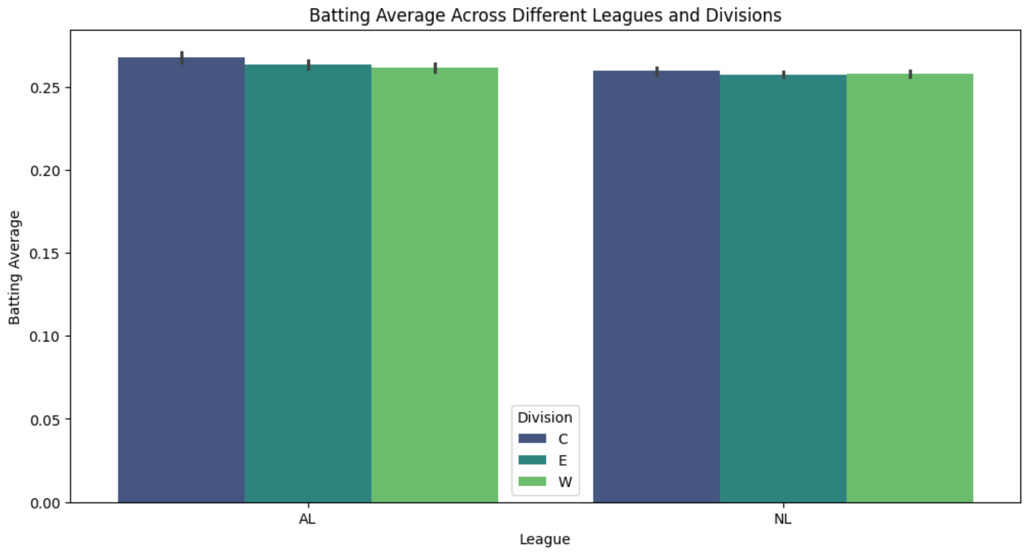
With this data, I was able to figure out which of the factors (Runs, Runs Against, Hits, At Bats, Earned Run Average, Fielding Percentage) correlate with the team's division rank the most to see the impacts that they have on overall performance.

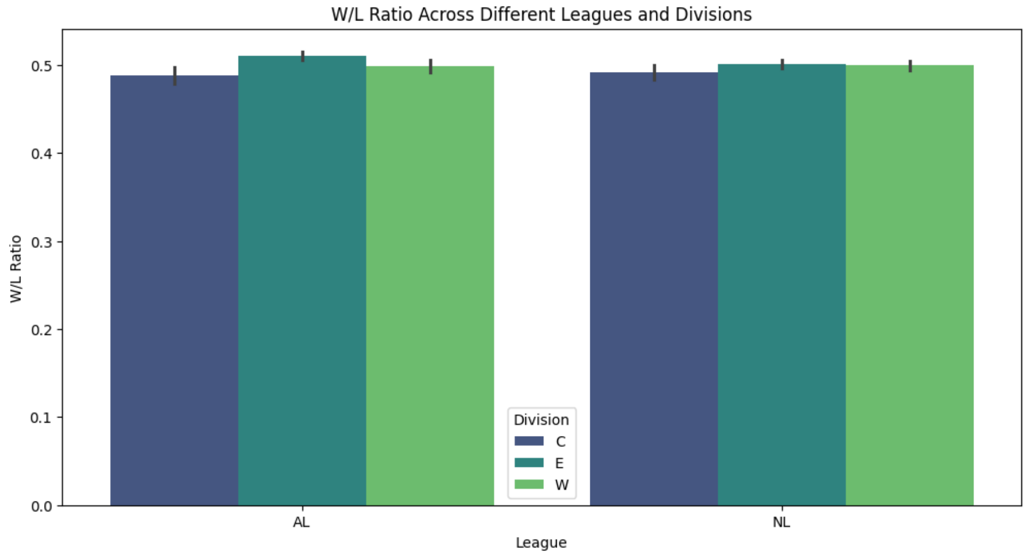


(Source: author)

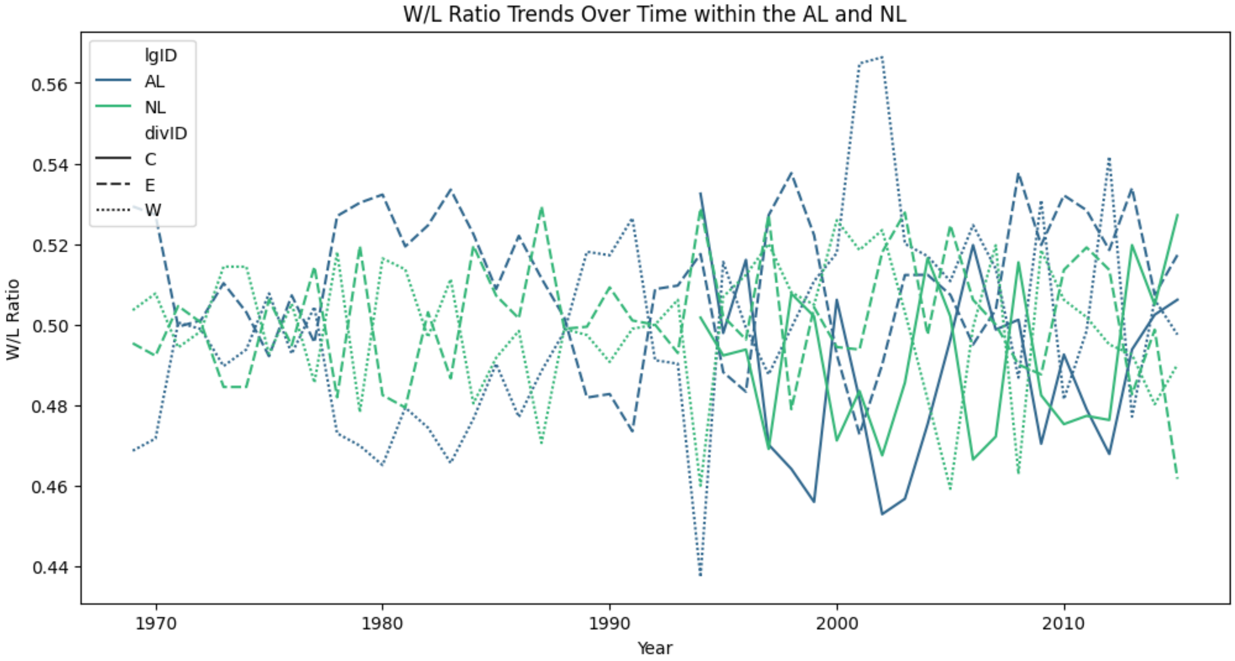
Based on the correlation values, we can see that runs and batting average (Hits / At Bats) are the most important factors when seeking out the lowest rank. (Lower rank number = Higher in standings.) Additionally, higher ERA and higher runs allowed are positively correlated with a higher rank. From this, we can infer that having strong pitching and overall defense are factors that contribute most to a team's success.

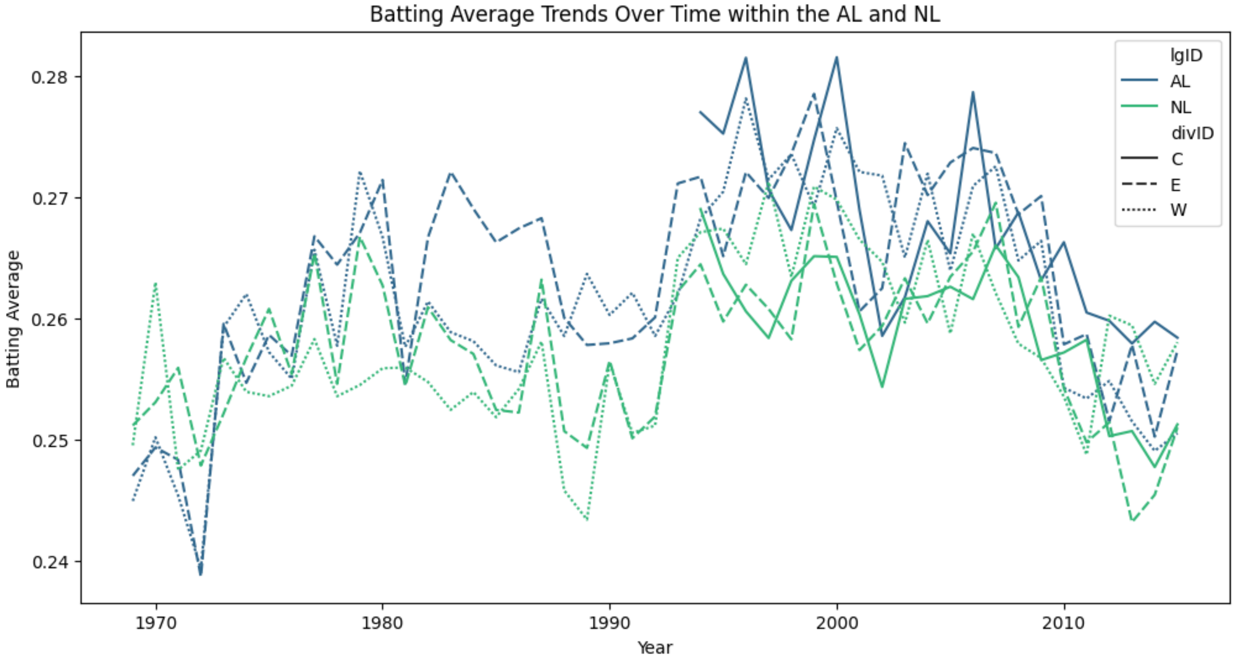
The next question that I set out to answer was if there was any variance in performance across different leagues and divisions. To do this, I summarized key metrics by division to see if one was stronger than others.





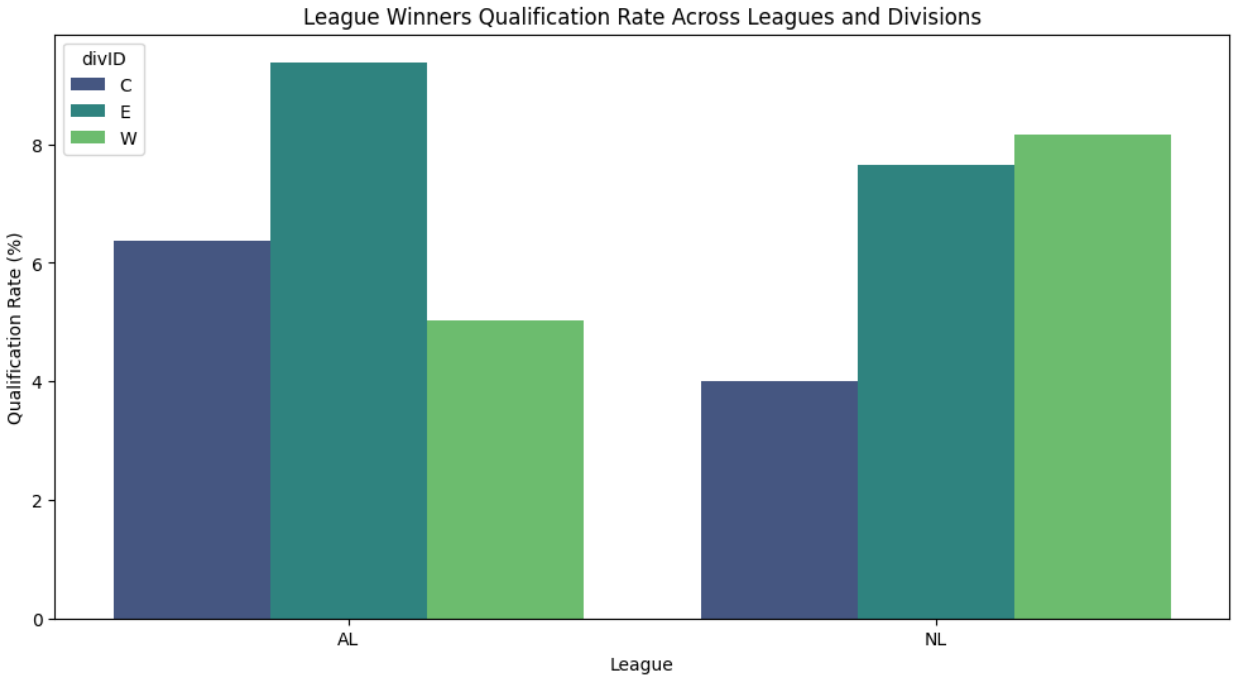
Based on this data, there are no consistent, clear differences between leagues and divisions. Because of this, I decided to dig further and investigate if there have been any trends in performance over time.

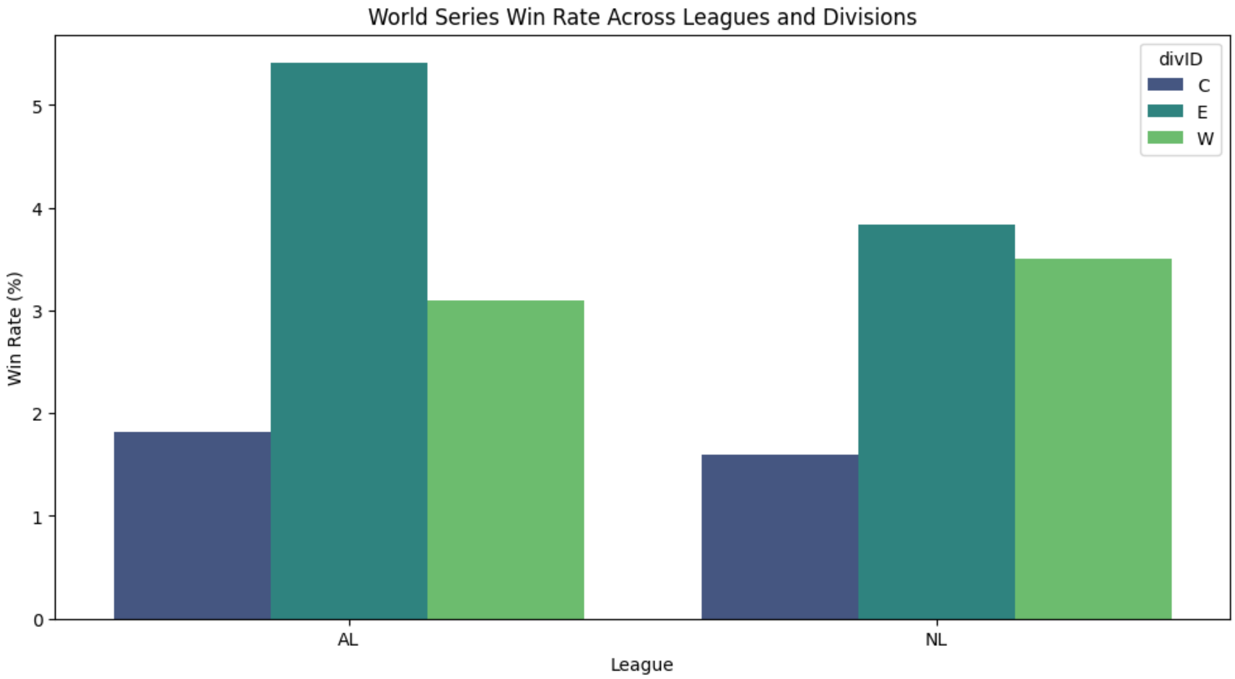




When looking at the Win/Loss ratio over time, there are no clear trends that we can decipher besides some years with clear outliers. The American League has had a historically higher batting average compared to the National League. This can be attributed to the fact that the American League has used the ‘Designated Hitter’ rule since 1973, where it was only adopted by the National League in 2022. The Designated Hitter rule is when a player bats in place of another player who would’ve originally batted. Usually this is a pitcher, as pitchers are historically not good hitters. As a result, it makes sense that there is an increase in the batting average for a league that has less statistically worse players at batting.

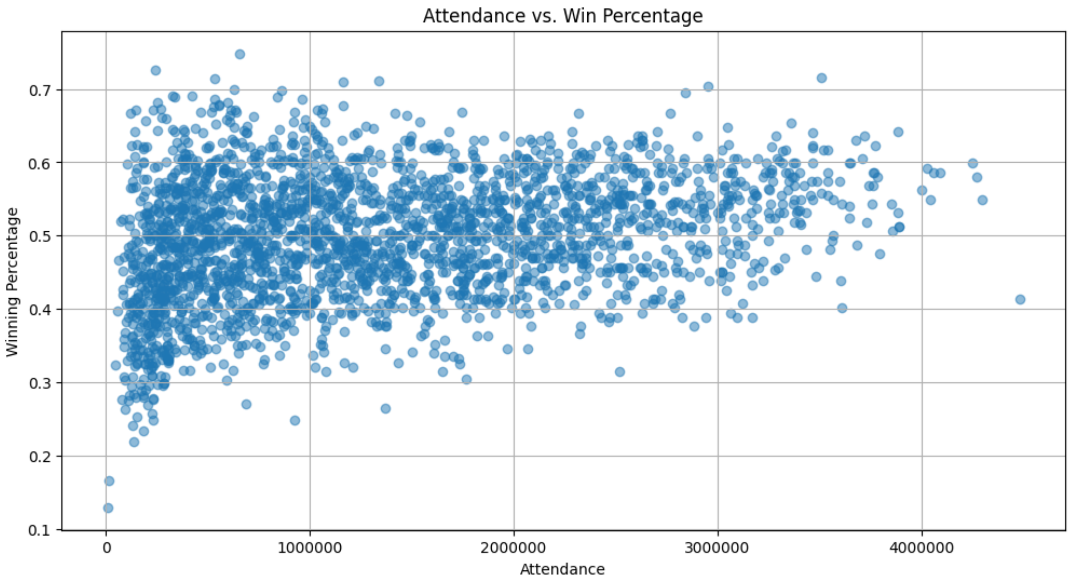
I was then interested in looking at any trends for who wins the league championship and for who wins the World Series.



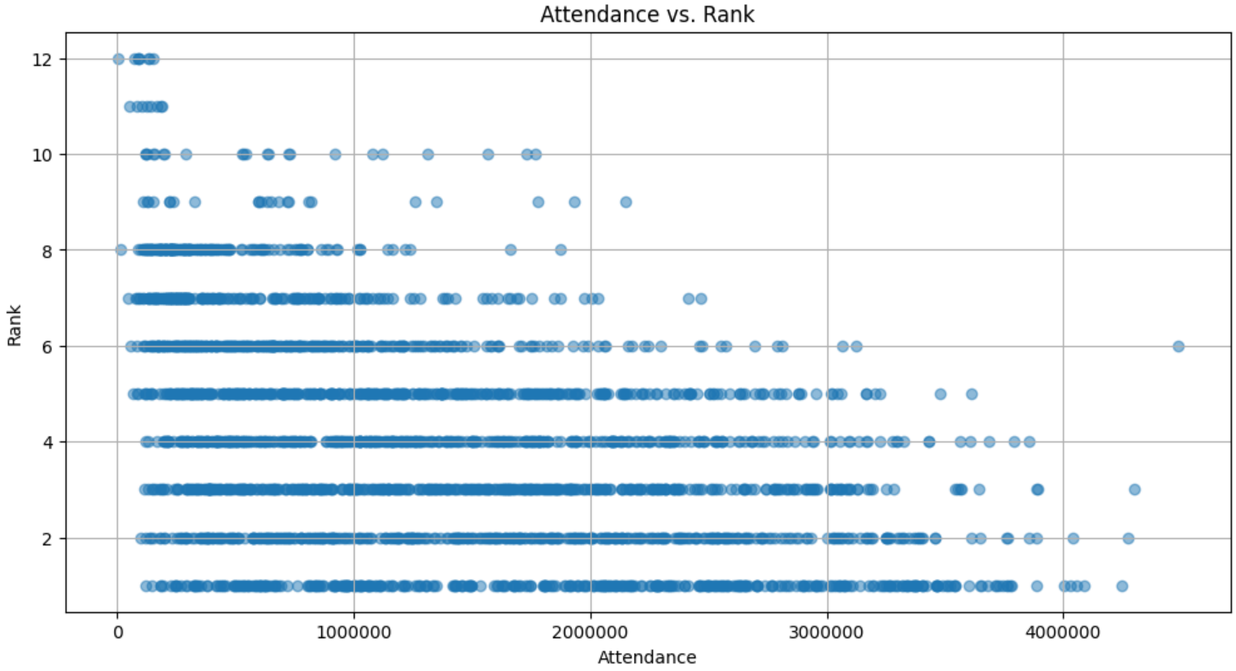


When looking at league winners, I found it most surprising that the East division wins the American League 9.4% of the time. This strong east division is also portrayed in the World Series win rate. In any given year, there is a 5.4% chance that the World Series winners will be a team from the East division of the American League. This makes sense, as the teams in this division are very strong, boasting the likes of the New York Yankees, Boston Red Sox, and Tampa Bay Rays.

The next research question that I wanted to ask was if there was a correlation between attendance and team performance.

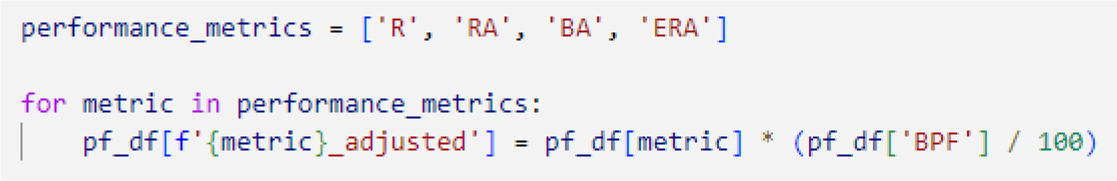


As we can see from the graph, there is a weak positive correlation, suggesting that teams don't win more just because the attendance increases. Testing the correlation between a team's attendance and their position in the division proves more useful.



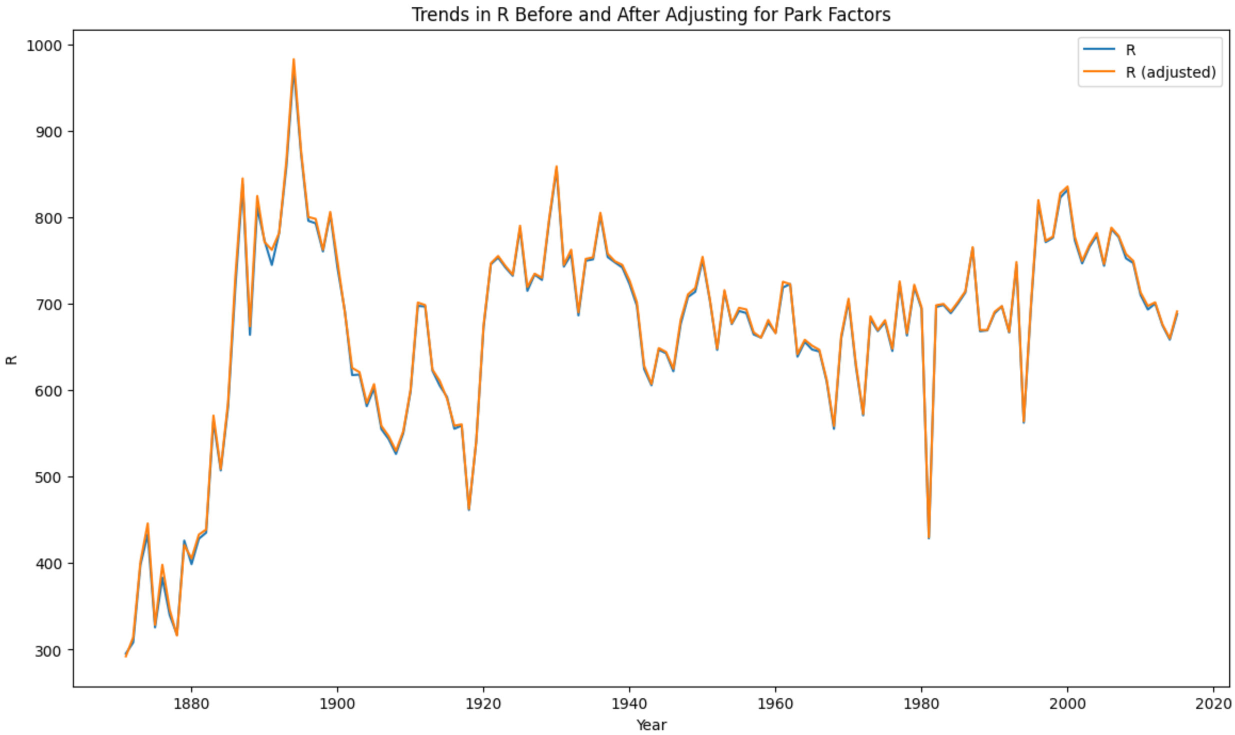
This graph shows a moderate negative correlation, suggesting that as teams climb in the standings (rank gets lower), the overall attendance tends to increase. This can likely be attributed to winning teams stirring up more interest for the fans and for the general population to tune in. Outliers in this data can be a result of a team's overall popularity, market size, stadium capacity, and other external factors.

The next question I wanted to answer was how park factors influenced team performance over time. Park factors (batting park factor and pitching park factor) are the three-year averages that compare the rate of stats for games at home versus the rate of stats for games when a team is on the road. If the park factor is higher than 100, that means that the batters/pitchers perform better at home versus on the road.

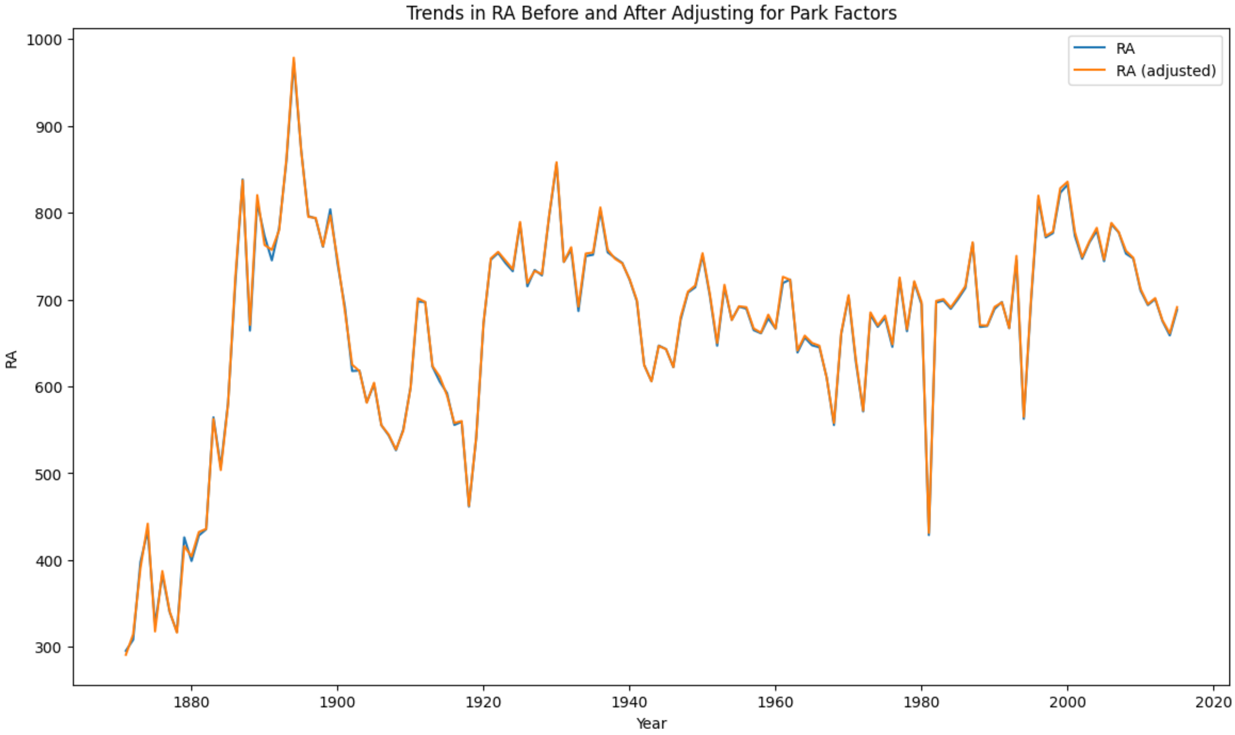


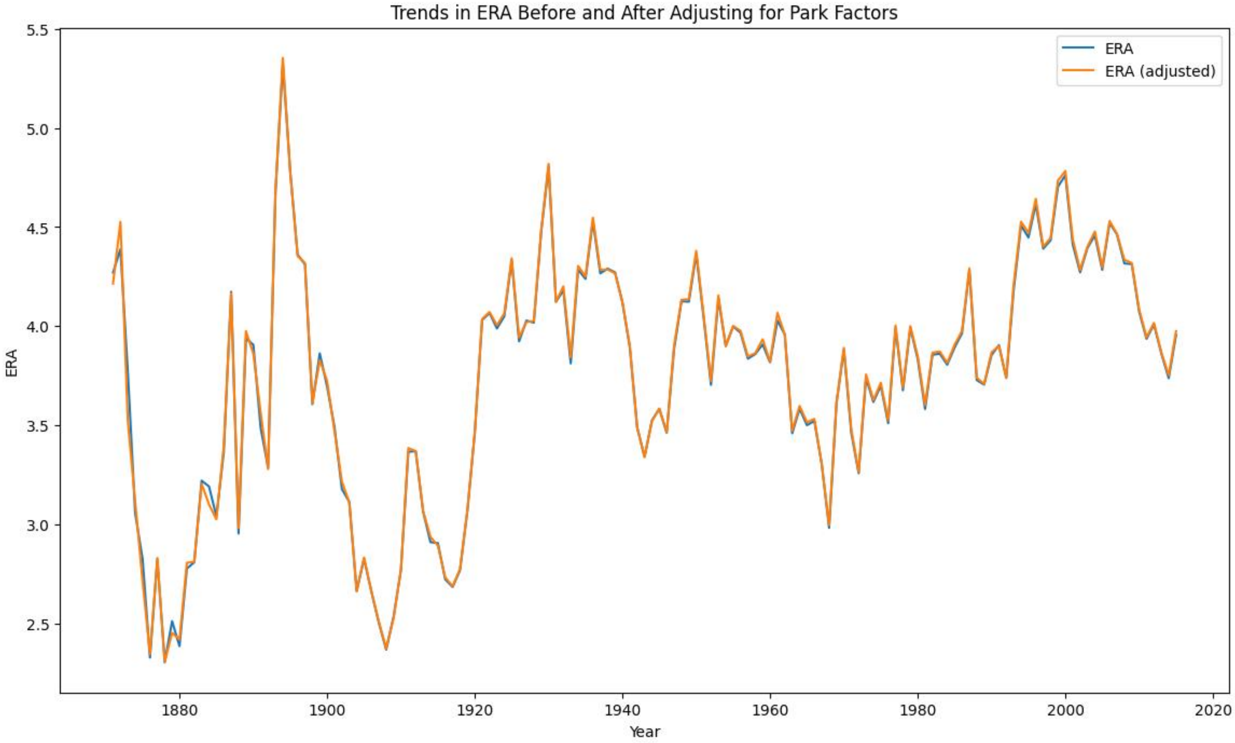
```
performance_metrics = ['R', 'RA', 'BA', 'ERA']

for metric in performance_metrics:
    pf_df[f'{metric}_adjusted'] = pf_df[metric] * (pf_df['BPF'] / 100)
```

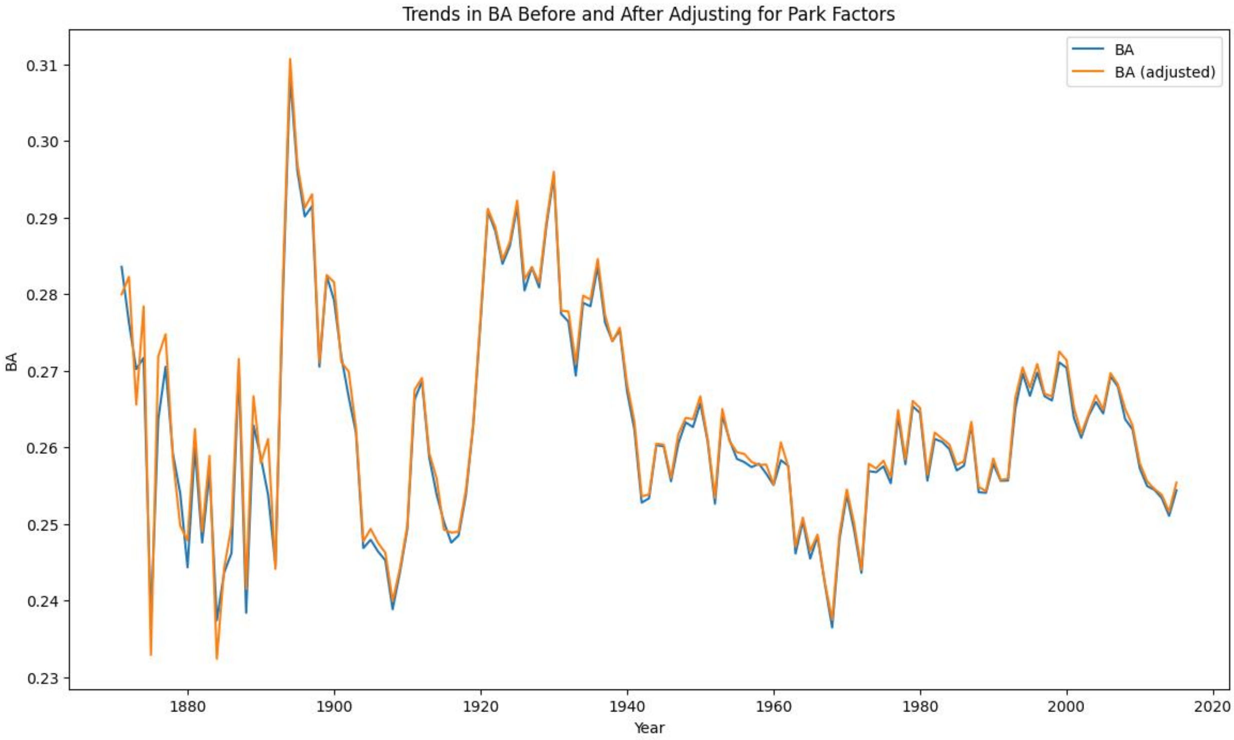
To do this, I adjusted each performance metric based on the batting park factor to get the adjusted metric values. With this, I plotted a line graph for each metric giving us a visual for the differences in values.



Trends in R Before and After Adjusting for Park Factors


Trends in RA Before and After Adjusting for Park Factors
RA
RA (adjusted)
1000
900
800
700
600
500
400
300
RA
1880
1900
1920
1940
1960
1980
2000
2020
Year

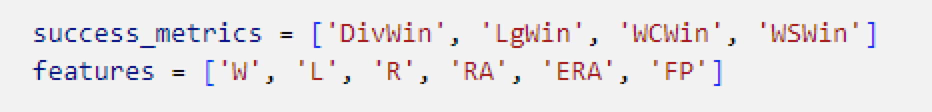

Trends in ERA Before and After Adjusting for Park Factors
ERA
ERA (adjusted)
5.5
5.0
4.5
4.0
3.5
3.0
2.5
ERA
1880
1900
1920
1940
1960
1980
2000
2020
Year



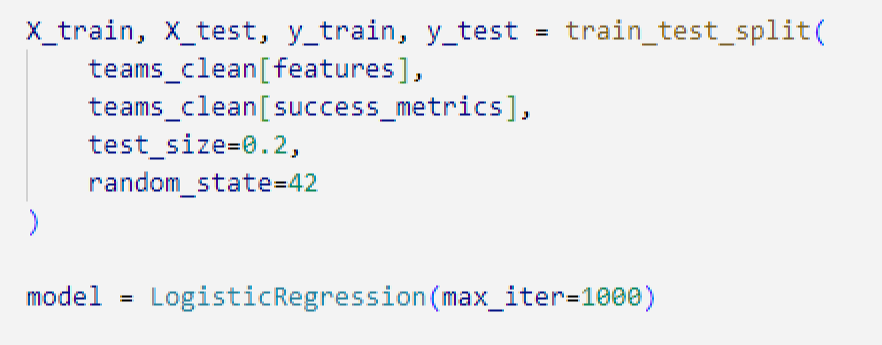
The only noticeable change in performance is seen in the Batting Average graph. After adjusting for park factors, we can see a slight increase in the overall batting average. This means that batters typically perform better when playing at home than they do when playing on the road. None of the other performance metrics show any measurable or significant increases in team performance.

The final research question that I set out to answer was whether or not we can predict a team's playoff success based on their regular season performance. To achieve this, I wanted to utilize a machine learning model. I decided on logistic regression as the model due to its ability to estimate the probability of an event occurring given a set of variables. A team's playoff success was attributed to whether or not they won their division. While not ideal, this was a limitation of the data that I was working with. The features of the model include Wins, Losses, Runs, Runs

Allowed, ERA, and Fielding Percentage. Based on these factors, I am hoping to be able to predict whether or not a team wins their division.

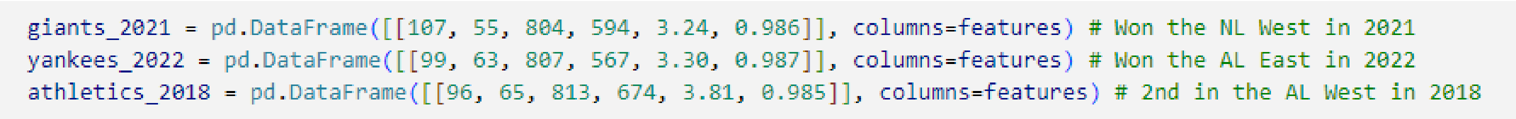


```
success_metrics = ['DivWin', 'LgWin', 'WCWin', 'WSWin']
features = ['W', 'L', 'R', 'RA', 'ERA', 'FP']
```



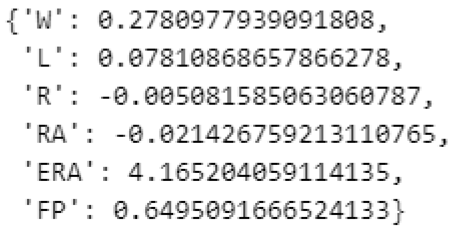
```
X_train, X_test, y_train, y_test = train_test_split(
    teams_clean[features],
    teams_clean[success_metrics],
    test_size=0.2,
    random_state=42
)

model = LogisticRegression(max_iter=1000)
```

After training and testing my model, I can validate it by testing against real statistics that are outside the bounds of the dataset to see how well it performs. I took three different teams from three different MLB seasons with different end rankings to see how well the model performed.

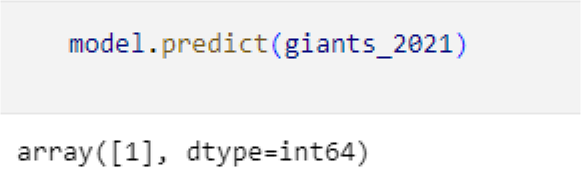


```
giants_2021 = pd.DataFrame([[107, 55, 804, 594, 3.24, 0.986]], columns=features) # Won the NL West in 2021
yankees_2022 = pd.DataFrame([[99, 63, 807, 567, 3.30, 0.987]], columns=features) # Won the AL East in 2022
athletics_2018 = pd.DataFrame([[96, 65, 813, 674, 3.81, 0.985]], columns=features) # 2nd in the AL West in 2018
```

When testing our model, we could expect the results to show a 1 for the first two predictions and a 0 for the final prediction (two division winners and one non-division winner.)

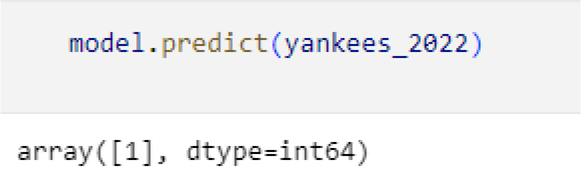


```
{'W': 0.2780977939091808,
 'L': 0.07810868657866278,
 'R': -0.005081585063060787,
 'RA': -0.021426759213110765,
 'ERA': 4.165204059114135,
 'FP': 0.6495091666524133}
```



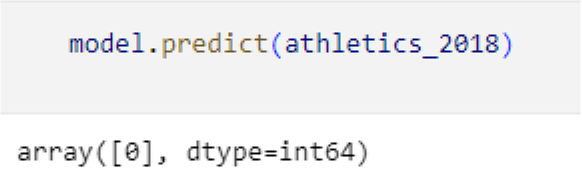
```
model.predict(giants_2021)
```

```
array([1], dtype=int64)
```



```
model.predict(yankees_2022)
```

```
array([1], dtype=int64)
```



```
model.predict(athletics_2018)
```

```
array([0], dtype=int64)
```

All three of our predictions are correct, further validating our model's accuracy. Based on the feature importance, we can see that win percentage, fielding percentage, and pitching performance (ERA and runs allowed) are the most important factors determining a team's playoff success based on regular season performance.

Conclusion

While professional baseball has changed quite a bit over the years, its continued popularity and success is a testament to its appeal to capture the hearts of fans all throughout the world. In this report, we looked at the factors that contribute to a team's success, if team performance varies between divisions, how big of a role attendance and park factors play for teams and predicting a team's playoff success based on their regular season performance. As professional baseball continues to evolve and develop, I hope my analysis can serve as a foundational reference for understanding the trends that have developed as well as its historical context.

References

Open Source Sports. (2019, November). *Baseball Databank*. Kaggle. https://www.kaggle.com/datasets/open-source-sports/baseball-databank/.

Team Chronology. Retrosheet. (n.d.). https://www.retrosheet.org/chronology.htm.