Name: Erkang Chen **USCID:** 8906487291

DSCI 510 Final Project

Abstract: This project aims to analyze the performance of the Los Angeles Dodgers players during the 2022 MLB season by utilizing three different data sources. The objective is to gain insights into individual player performance and team dynamics that contributed to the Dodgers' success during the regular season. The data sources include publicly available SportsDataIO API, and scraping data from various sources. Through data cleaning, exploratory data analysis, and machine learning models, the project aims to uncover patterns and trends that can provide insights into the factors that led to the team's success. The results of this project could be useful in identifying key areas for improvement in the team's performance in future seasons.

Motivation: The motivation for this project is to gain a better understanding of the performance of the Los Angeles Dodgers players during the 2022 MLB season and how it contributes to the team's success. By utilizing various data sources, including traditional and advanced baseball statistics, the aim is to analyze player performance, compare it with other teams and players in the league, and explore the consistency of the Dodgers' performance over the last decade. The hypothesis is that the performance of individual players will be crucial to the team's overall success, and advanced metrics will provide a more comprehensive assessment of player performance than traditional statistics. Through this project, we can also identify any potential areas of improvement for the team and individual players. By analyzing the team's performance over multiple seasons, we can explore the consistency of the Dodgers and their ability to maintain a high level of performance over time.

Three Data Sources:

Name: Erkang Chen **USCID:** 8906487291

1. **Major League baseball players batting/pitching stats by teams in 2022 season from baseball-reference.com**

The first data source for this project is Major League Baseball players' batting statistics by teams in the 2022 season, which will be scraped from baseball-reference.com. The data will be extracted using a specific URL structure (https://www.baseball-reference.com/teams/{teamname}/2022.shtml) that corresponds to each team in the league. The data will provide valuable insights into the performance of individual players, such as batting average, home runs, RBIs, and stolen bases, and will be used to analyze the Los Angeles Dodgers players' performance in particular. In addition to batting statistics, the data source from baseball-reference.com will also provide valuable pitching data. This data will include the number of innings pitched, earned run average (ERA), strikeouts, walks, and hits allowed, among other metrics. The pitching data will give insights into how well a pitcher is performing and will be crucial in analyzing the Los Angeles Dodgers' overall performance.

	Pos	Name	Age	G	PA	AB	R	H	2B	3B	...	OBP	SLG	OPS	OPS+	TB	GDP	HBP	SH	SF	IBB
0	C	Will Smith	27	137	578	508	68	132	26	3	...	.343	.465	.807	121	236	11	10	0	4	4
1	1B	Freddie Freeman*	32	159	708	612	117	199	47	2	...	.407	.511	.918	153	313	6	5	0	7	12
2	2B	Gavin Lux*	24	129	471	421	66	116	20	7	...	.346	.399	.745	106	168	3	0	0	3	0
3	SS	Trea Turner	29	160	708	652	101	194	39	4	...	.343	.466	.809	122	304	9	3	0	6	1
4	3B	Max Muncy*	31	136	565	464	69	91	22	1	...	.329	.384	.713	97	178	2	5	0	6	1

	Pos	Name	Age	W	L	W-L%	ERA	G	GS	GF	...	WP	BF	ERA+	FIP	WHIP	H9	HR9	BB9	SO9	SO/W
0	SP	Tyler Anderson*	32	15	5	.750	2.57	30	28	0	...	6	707	161	3.31	1.002	7.3	0.7	1.7	7.0	4.06
1	SP	Julio Urías*	25	17	7	.708	2.16	31	31	0	...	0	689	192	3.71	0.960	6.5	1.2	2.1	8.5	4.05
2	SP	Tony Gonsolin	28	16	1	.941	2.14	24	24	0	...	5	498	194	3.28	0.875	5.5	0.8	2.4	8.2	3.40
3	SP	Clayton Kershaw*	34	12	3	.800	2.28	22	22	0	...	0	493	182	2.57	0.942	6.8	0.7	1.6	9.8	5.96
4	SP	Andrew Heaney*	31	4	4	.500	3.10	16	14	1	...	0	310	134	3.75	1.087	7.4	1.7	2.4	13.6	5.79

2. **League batting & pitching by teams through 2012-2022 seasons (API)**

Name: Erkang Chen **USCID:** 8906487291

The second data source for this project will be fetched from an external public API called the SportsDataIO API (https://api.sportsdata.io/v3/mlb/scores/json/TeamSeasonStats/{year}?key=eb1aa7acf352437693545e77c66c2e5a). This API will provide two datasets: League batting by teams through 2012-2022 seasons and League pitching by teams through 2012-2022 seasons. These datasets will include data on all Major League Baseball teams over the past ten years, providing valuable insights into how teams' performance has changed over time. The data will enable us to evaluate the performance of the Los Angeles Dodgers as a team, including their batting and pitching performance, and compare it to that of other teams in the league.

	Team	At bats	Wins	Runs	Hits	Singles	Doubles	Triples	Home runs	Stolen bases	Runs batted in	Outs	Strikeouts	Walks	Year
1	ARI	7543.0	111.9	1013.7	1955.5	1258.1	424.0	45.6	227.9	128.4	980.5	5587.5	1748.3	744.4	2012
2	WSH	7754.3	135.3	1009.5	2027.3	1309.2	415.7	34.5	267.9	145.0	950.1	5727.0	1829.8	661.5	2012
3	TOR	7577.5	100.8	988.8	1858.8	1213.9	341.1	30.4	273.4	169.9	934.9	5718.7	1727.6	653.2	2012
4	TEX	7719.8	128.4	1115.8	2106.0	1367.2	418.4	44.2	276.2	125.7	1075.8	5613.8	1523.2	660.1	2012
5	TB	7454.6	124.3	962.6	1785.6	1157.3	345.2	41.4	241.7	185.1	918.4	5669.0	1827.1	788.6	2012

	Team	Innings pitched	Earned runs	Hits allowed	Wins	Walks allowed	Strikeouts pitched	Year
1	ARI	1979.0	867.3	1977.6	111.9	5.0	1657.2	2012
2	WSH	2027.3	752.6	1789.8	135.3	5.8	1829.8	2012
3	TOR	1992.8	1028.8	1985.9	100.8	6.8	1578.5	2012
4	TEX	1991.4	889.4	1903.0	128.4	5.3	1776.0	2012
5	TB	2014.9	715.4	1702.8	124.3	5.5	1909.9	2012

3. **Fangraph.com (getting advancing performance data for players)**

The third data source for this project will be scraped from Fangraphs.com(https://www.fangraphs.com/leaders.aspx?pos=all&stats=bat&lg=all&qual=0&type=8&season=2022&month=0&season1=2022&ind=0&team=0&rost=0&age=0&filter=&players=0&startdate=&enddate=&page={}), a site that provides advanced baseball statistics that can help provide a more granular assessment of player

performance than traditional statistics. The data will include advanced metrics for batters and pitchers such as Wins Above Replacement (WAR), Fielding Independent Pitching (FIP), and more.

Rank	Name	Team	G	PA	HR	R	RBI	SB	BB%	K%	...	AVG	OBP	SLG	wOBA	xwOBA	wRC+	BsR	Off	Def	WAR
1	Aaron Judge	NYY	157	696	62	133	131	16	15.9%	25.1%	...	.311	.425	.686	.458	.463	207	2.1	86.1	1.1	11.5
2	Manny Machado	SDP	150	644	32	100	102	9	9.8%	20.7%	...	.298	.366	.531	.382	.338	152	3.0	41.9	6.9	7.4
3	Nolan Arenado	STL	148	620	30	73	103	5	8.4%	11.6%	...	.293	.358	.533	.381	.339	151	-1.6	35.3	13.4	7.3
4	Paul Goldschmidt	STL	151	651	35	106	115	7	12.1%	21.7%	...	.317	.404	.578	.419	.367	177	3.0	61.6	-15.8	7.1
5	Freddie Freeman	LAD	159	708	21	117	100	13	11.9%	14.4%	...	.325	.407	.511	.393	.403	157	5.4	52.6	-8.8	7.1

Rank	Name	Team	W	L	SV	G	GS	IP	K/9	BB/9	...	BIP	LOB%	GB%	HR/FB	vFA (pi)	ERA	xERA	FIP	xFIP	WAR
1	Aaron Nola	PHI	11	13	0	32	32	205.0	10.32	1.27	...	.289	73.0%	43.6%	9.8%	92.9	3.25	2.74	2.58	2.77	6.3
2	Carlos Rodon	SFG	14	8	0	31	31	178.0	11.98	2.63	...	.293	75.1%	34.1%	6.5%	95.5	2.88	2.64	2.25	2.91	6.2
3	Justin Verlander	HOU	18	4	0	28	28	175.0	9.51	1.49	...	.240	80.5%	37.9%	6.2%	95.1	1.75	2.66	2.49	3.23	6.1
4	Sandy Alcantara	MIA	14	9	0	32	32	228.2	8.15	1.97	...	.262	78.8%	53.4%	8.5%	98.0	2.28	2.92	2.99	3.29	5.7
5	Kevin Gausman	TOR	12	10	0	31	31	174.2	10.56	1.44	...	.363	74.0%	39.2%	8.5%	94.9	3.35	3.34	2.38	2.75	5.7

Data derived from combining (3) above for analysis:

This part will include in the analysis

Graphs, pictures, tables, visualizations:

This part will include in the analysis

Technical solution:

To analyze the Los Angeles Dodgers performance in the 2022 season, the first step is to define the variables of interest, such as runs scored, batting average, on-base percentage, earned run average, and win-loss record. Next, data is scraped from fangraphs.com and baseball-reference.com using web scraping tools like Beautiful Soup or Scrapy, for both team-level data and player-level data. Also, one data set is

fetched from an external public API called the SportsDataIO API. The scraped data is then cleaned and merged based on common identifiers. The next step is to choose an appropriate model, such as decision trees or random forests, or linear regression for continuous data analysis. The data is split into training and testing sets, and the model is trained using the training data and fitted to the testing data to evaluate accuracy and performance. Finally, analysis is performed on the model results to gain insights into the team's performance, players' performance, and players' contribution.

I have faced some challenges in doing this project, such as merging datasets with different formats and selecting the most suitable model for the given data. To tackle these challenges, data cleaning and experimenting with different model types may be necessary to find the best fit. Another challenge you might encounter is finding the right API that provides the required data, which may involve searching through multiple APIs and documentation. Furthermore, considering the cost and availability of different API options is essential since not all APIs are free or offer the same level of access. Once you have the data, determining the most appropriate model for analyzing it can be challenging and may require experimentation and testing to find the best results. It's crucial to understand the strengths and limitations of different models to select the most suitable one for your data. For the APi part, I found several API and finally I got one I can use. For the model selection, I try to use the data to fit into different models and choose the best model that can describe and analyze the data.

Analysis, insights:

My Analysis focus on answering four questions:

- Which players had the greatest impact on the team's success?

- How does the team's performance compare to other teams in the league?
- How is the consistency of the Los Angeles Dodgers over the last 10 seasons?
- How does the performance of individual players relate to the overall success of the team?

By answering the first question I use datasets from baseball-reference.com for both batters and pitchers. I fit the data into a random regressor model and calculate the mean square error. By doing so, I can see the accuracy of my model. I use these data as X components ("G", "PA", "H", "2B", "3B", "HR", "BB","SO","BA")[1] and the Y component using "OPS" (a measure to measure the offensive performance of a batter).

[1] G (Games played): The number of games a player has appeared in during a season.
PA (Plate appearances): The number of times a player has come up to bat during a season. Plate appearances include official at-bats, walks, hit-by-pitches, and sacrifices.

H (Hits): The number of times a player has safely reached base by hitting a fair ball and reaching base without the help of an error or fielder's choice.

2B (Doubles): A hit where the batter reaches second base safely without the help of an error or fielder's choice.

3B (Triples): A hit where the batter reaches third base safely without the help of an error or fielder's choice.

HR (Home runs): A hit where the batter hits the ball over the outfield wall, resulting in a run being scored.

BB (Walks): When a pitcher throws four balls outside the strike zone, allowing the batter to take first base.

SO (Strikeouts): When a batter swings at and misses three pitches or is called out by the umpire after failing to swing at three pitches in the strike zone.

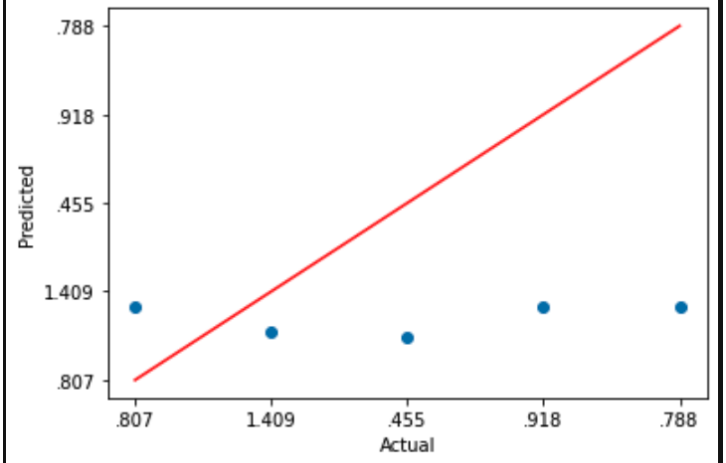
BA (Batting average): A statistic that measures a batter's success at getting a hit. It is calculated by dividing the number of hits by the number of at-bats.

OPS (On-base plus Slugging) : A statistic used in baseball to evaluate a player's overall offensive performance

Name: Erkang Chen **USCID:** 8906487291

Then I got this

```
Mean Squared Error: 0.14923204481999994
```



It seems like the model is not too bad so I performed the RandomForestRegressor to find which X components impact the “OPS” most. Here is the result.

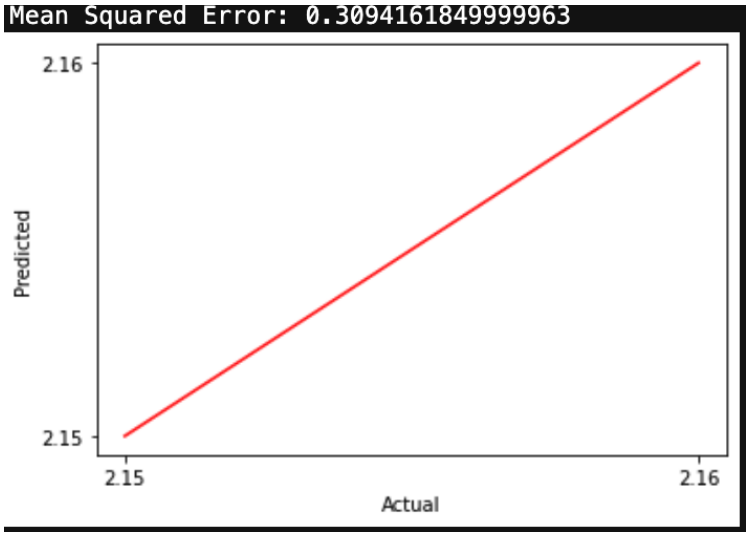
```
Feature ranking:
1. BB (0.282673)
2. HR (0.191576)
3. G (0.133917)
4. BA (0.126580)
5. PA (0.123256)
6. SO (0.063001)
7. H (0.056651)
8. 2B (0.011609)
9. 3B (0.010737)
```

Then I times the players data with the weights of each component and I add them together to get a new score for each player. Then I rank the players’ performance based on the score I got. **So for batters, players who had the greatest impact on the team's success are Freddie Freeman, Trea Turner, Mookie Betts, Max Muncy, and Will Smith.**

	Name	G	PA	H	2B	3B	HR	BB	SO	BA	HR_score	BA_score	BB_score	PA_score	H_score	G_score	Total_score
0	Freddie Freeman*	159	708	199	47	2	21	84	102	0.325	4.02	0.041139	23.74	87.27	11.27	21.29	147.60
1	Trea Turner	160	708	194	39	4	21	45	131	0.298	4.02	0.037721	12.72	87.27	10.99	21.43	136.43
2	Mookie Betts	142	639	154	40	3	35	55	104	0.269	6.71	0.034050	15.55	78.76	8.72	19.02	128.75
3	Max Muncy*	136	565	91	22	1	21	90	141	0.196	4.02	0.024810	25.44	69.64	5.16	18.21	122.47
4	Will Smith	137	578	132	26	3	24	56	96	0.260	4.60	0.032911	15.83	71.24	7.48	18.35	117.49

Name: Erkang Chen **USCID:** 8906487291

For analyzing pitchers performance, I did the same thing, with different X components and Y components, since there are different criteria between pitchers and batters. The X components are "W", "IP", "BB", "SO", "ER", "HR", "BF". Y components are "ERA"[2]. For Y component “ERA”, the lower ERA a pitcher has, the better pitching performance the pitcher has. I run the RandomForestRegressor model for pitchers, calculating the mean square error and get the weights for each X components, calculating the scores for pitchers.



[2]W: Short for "wins," this statistic measures the number of games a pitcher has won.

IP: Short for "innings pitched," this statistic measures the number of innings a pitcher has thrown.

BB: Short for "base on balls," this statistic measures the number of times a pitcher has given a batter a free pass to first base by throwing four balls.

SO: Short for "strikeouts," this statistic measures the number of times a pitcher has struck out a batter.

ER: Short for "earned runs," this statistic measures the number of runs that were scored against a pitcher that were not the result of an error by the fielding team.

HR: Short for "home runs," this statistic measures the number of times a batter has hit the ball out of the park.

BF: Short for "batters faced," this statistic measures the number of batters that a pitcher has faced during a game or season.

ERA: Short for "earned run average," this statistic measures the number of earned runs a pitcher gives up per nine innings pitched. It is calculated by taking the total number of earned runs a pitcher has given up and dividing it by the number of innings pitched, then multiplying by nine.

Name: Erkang Chen **USCID:** 8906487291

```
Feature ranking:
1. W (0.302221)
2. SO (0.154164)
3. IP (0.139913)
4. BB (0.116023)
5. ER (0.110453)
6. HR (0.107627)
7. BF (0.069599)
```

	Name	W	IP	BB	SO	ER	HR	BF	W_score	SO_score	IP_score	BB_score	ER_score	HR_score	BF_score	Total_score
0	Tyler Anderson*	15	178.2	34	138	51	14	707	4.53	21.274632	24.93	3.94	5.63	1.51	49.21	89.756968
1	Julio Urías*	17	175.0	41	166	42	23	689	5.14	25.591224	24.48	4.76	4.64	2.48	47.95	89.447633
2	Tony Gonsolin	16	130.1	35	119	31	11	498	4.84	18.345516	18.20	4.06	3.42	1.18	34.66	66.367264
3	Clayton Kershaw*	12	126.1	23	137	32	10	493	3.63	21.120468	17.64	2.67	3.53	1.08	34.31	62.861283

I select the top 4 pitchers who have the highest scores. **For pitchers, players who had the greatest impact on the team's success are Tyler Anderson, Julio Urias, Tony Gonsolin, and Clayton Kershaw.**

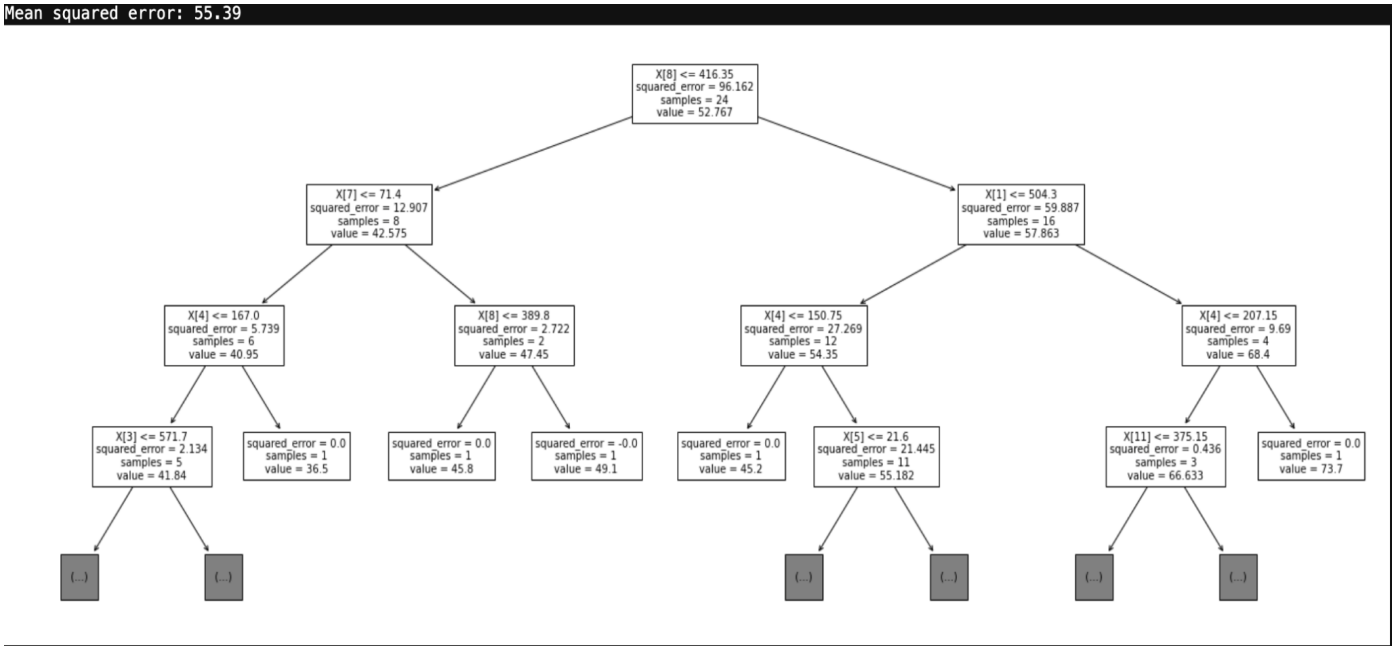
To Answer the second and third questions:

How does the team's performance compare to other teams in the league?

How is the consistency of the Los Angeles Dodgers over the last 10 seasons?

I use data fetched from API, calculate the mean square error, and run them into a decision tree model. For batting stats I use those measures as X components ('At bats', 'Runs', 'Hits', 'Singles', 'Doubles', 'Triples', 'Home runs', 'Stolen bases', 'Runs batted in', 'Outs', 'Strikeouts', 'Walks'). Y component is “Wins” Then, I got this graph.

Name: Erkang Chen **USCID:** 8906487291



Since we have more features and the data numbers are relatively larger, the mean square error is still acceptable. From these decision trees we can select the most important features by looking at the nodes of the decision tree. The most important one is X[8], which is 'Runs batted in', the second import features are X[1] and X[7], which are “Runs” and “Stolen Bases”. After I know which batting statistics contribute mostly to wins. I extract the Los Angeles Dodgers “Runs”, “RBI”, and “Stolen Base” performance through the league:

	Team	At bats	Wins	Runs	Hits	Singles	Doubles	Triples	Home runs	Stolen bases	Runs batted in	Outs	Strikeouts	Walks	Year
0	LAD	3669.3	73.7	562.4	941.6	564.4	215.8	20.6	140.8	65.1	539.2	2727.7	912.3	403.0	2022

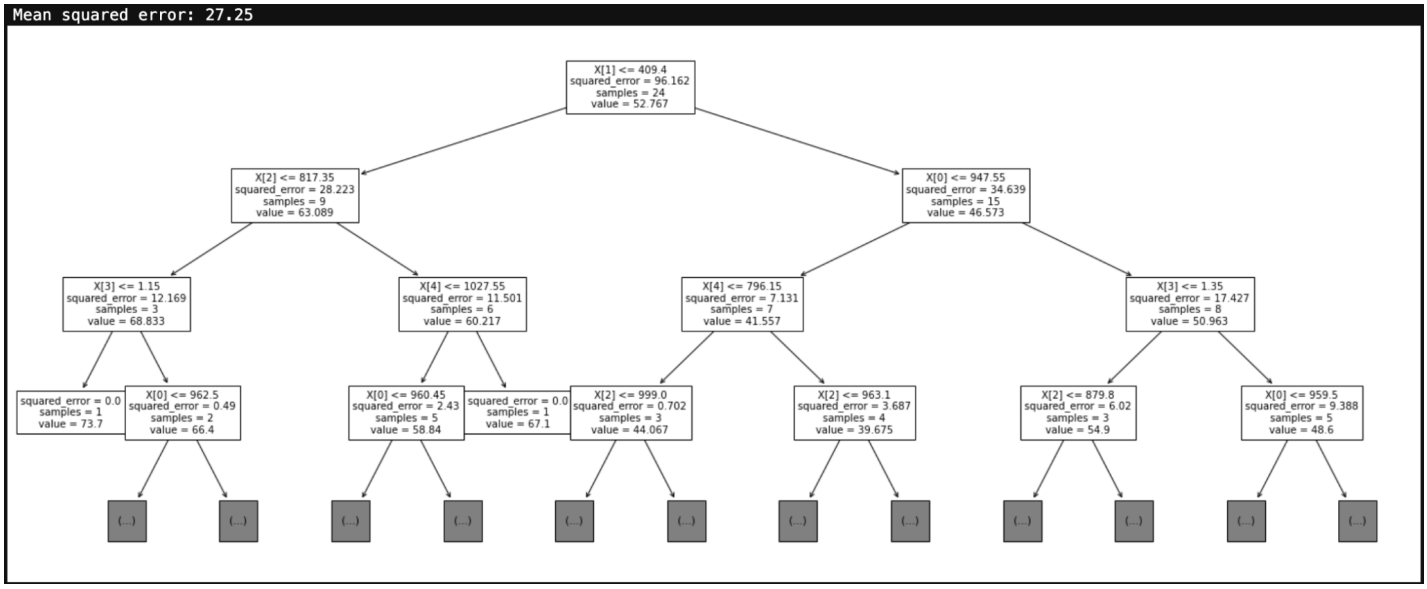
	Team	At bats	Wins	Runs	Hits	Singles	Doubles	Triples	Home runs	Stolen bases	Runs batted in	Outs	Strikeouts	Walks	Year
0	LAD	3669.3	73.7	562.4	941.6	564.4	215.8	20.6	140.8	65.1	539.2	2727.7	912.3	403.0	2022

	Team	At bats	Wins	Runs	Hits	Singles	Doubles	Triples	Home runs	Stolen bases	Runs batted in	Outs	Strikeouts	Walks	Year
8	LAD	3669.3	73.7	562.4	941.6	564.4	215.8	20.6	140.8	65.1	539.2	2727.7	912.3	403.0	2022

I can see the Runs and RBI are the number 1 of the league, and stolen bases are number 9 of the league. **Therefore, the batting performance is definitely top of the league.**

For pitching Performance I use 'Innings pitched','Earned runs','Hits allowed','Walks allowed','Strikeouts pitched' as X components and ‘Wins’ as Y component. Here is my model performance.

Name: Erkang Chen **USCID:** 8906487291



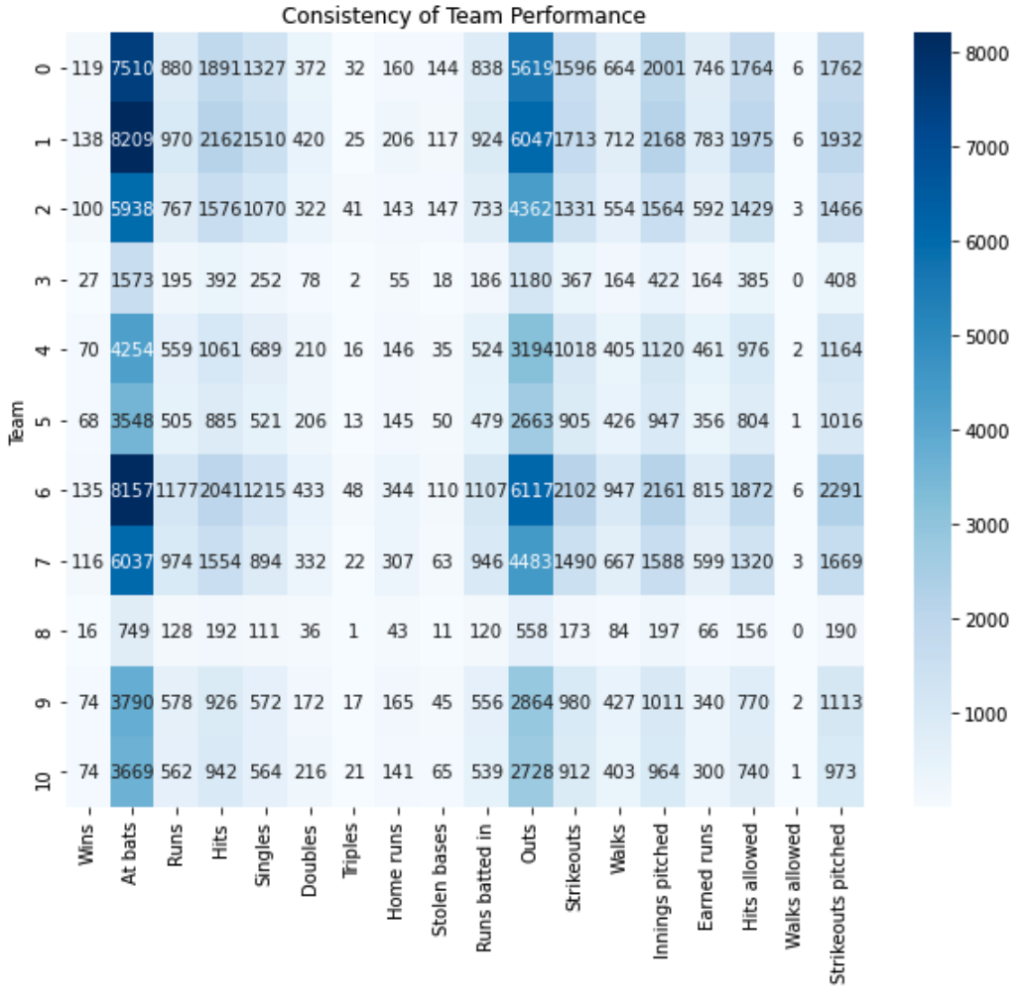
We can see the most important factor is x[1] which is “Earned runs” and the second important factors are x[2] and x[0] which are “Innings pitched” and “Hits allowed”.
I extract the Los Angeles Dodgers “Earned runs”,“Innings pitched”, and “Hits allowed” Performance through the league:

29	LAD	963.5	299.5	739.7	73.7	1.1	972.8	2022
2	LAD	963.5	299.5	739.7	73.7	1.1	972.8	2022
29	LAD	963.5	299.5	739.7	73.7	1.1	972.8	2022

By seeing the above data we can see that the Los Angeles Dodgers pitchers play second most innings and allow the least earned runs and hits allowed. (For hits allowed and earned runs, the less they are, the better the pitching performance, as it indicates that the pitcher was able to limit the number of hits and runs scored by the opposing team.) **Therefore, the pitching performance is the best in the league.**

To answer the consistency of the team in the last ten seasons I created a heatmap to see how it works, using the concatenation of the pitching data and batting data.

Name: Erkang Chen **USCID:** 8906487291



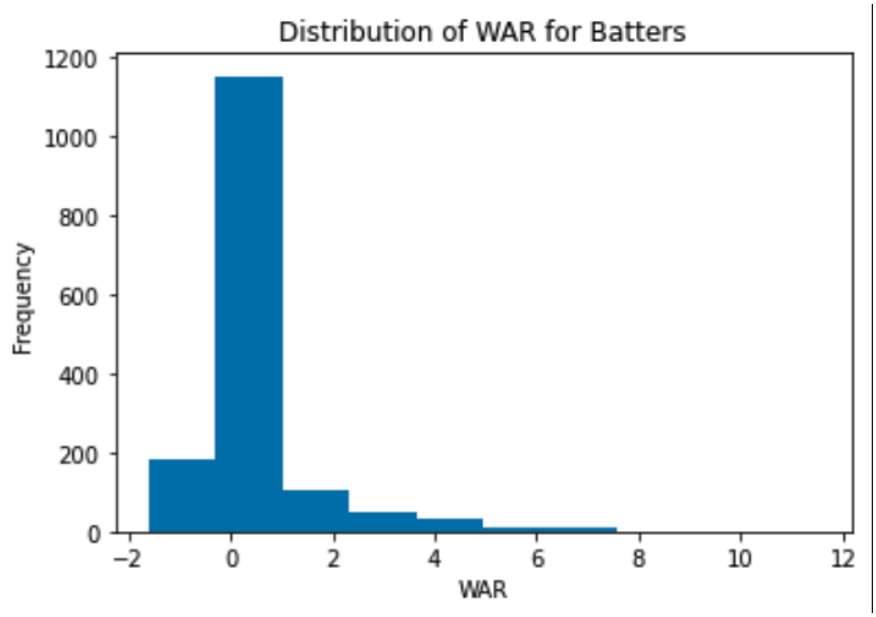
We can see from the color that only season 3 and season 8 (2014 season and 2020 season) did not perform well. Since 2020 Covid, MLB has had a much shorter season and Dodgers won the World Championship that year. Therefore, only the 2014 season is not on the top level. **Therefore, the Los Angeles Dodgers has been a consistent team in the last ten seasons.**

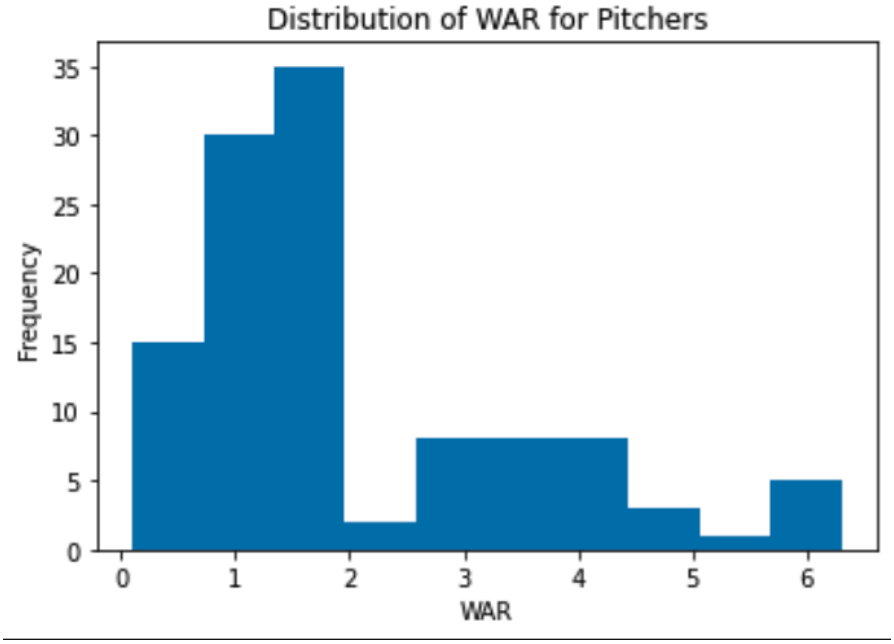
To answer the last question:

How does the performance of individual players relate to the overall success of the team?

We know that the team has pitching and batting. It is hard to measure how they contribute to the team. Now, we have a new measurement called “WAR”. WAR stands for "Wins Above Replacement," and it is a comprehensive statistic used in baseball to measure a player's overall value and contribution to his team. The higher WAR a player has, the more valuable they are to their team compared to a replacement-level player. I

have drawn two histograms to show the league WAR distribution for batters and pitchers.





Based on the graph I have selected 6 players, 4 batters and 2 pitchers who have high WARs.

	Name	Team	G	PA	HR	R	RBI	SB	BB%	K%	...	AVG	OBP	SLG	wOBA	xwOBA	wRC+	BsR	Off	Def	WAR
4	Freddie Freeman	LAD	159	708	21	117	100	13	11.9%	14.4%	...	0.325	0.407	0.511	0.393	0.403	157.0	5.4	52.6	-8.8	7.1
8	Mookie Betts	LAD	142	639	35	117	82	12	8.6%	16.3%	...	0.269	0.340	0.533	0.373	0.344	144.0	4.3	37.2	3.5	6.5
11	Trea Turner	LAD	160	708	21	101	100	27	6.4%	18.5%	...	0.298	0.343	0.466	0.350	0.335	128.0	6.5	29.5	7.1	6.3
98	Max Muncy	LAD	136	565	21	69	69	2	15.9%	25.0%	...	0.196	0.329	0.384	0.318	0.339	106.0	2.1	6.3	-2.8	2.4

Name: Erkang Chen **USCID:** 8906487291

	Name	Team	W	L	SV	G	GS	IP	K/9	BB/9	...	BIP	LOB%	GB%	HR/FB	vFA (pi)	ERA	xERA	FIP	xFIP	WAR
15	Tyler Anderson	LAD	15	5	0	30	28	178.2	6.95	1.71	...	0.256	77.8%	40.1%	6.4%	90.7	2.57	3.10	3.31	4.10	4.0
23	Julio Urias	LAD	17	7	0	31	31	175.0	8.54	2.11	...	0.229	86.6%	39.7%	10.8%	93.1	2.16	2.81	3.71	3.81	3.2

I would say that the above six players have contributed the most to the Los Angeles Dodgers 2022 season.

Conclusion:

I believe that the Los Angeles Dodgers' performance in the 2022 season is the best among all Major League Baseball teams, as demonstrated by both team and individual player statistics. Key players, such as Freddie Freeman, Mookie Betts, Trea Turner, Max Muncy, Tyler Anderson, and Julio Urias, have all contributed to the team's success. Not only did these players perform impressively, but the team as a whole displayed a high level of consistency throughout the last ten seasons. Such a high level of performance suggests that the team has a strong foundation for future success. However, it is important to keep in mind that baseball is an unpredictable sport, and future performance is never a guarantee.

limitations, challenges, future work ideas:

While the Los Angeles Dodgers demonstrated an impressive performance during the 2022 regular season, winning 111 games and ranking first in the league in multiple categories, they fell short in the playoffs and did not win the World Series championship. Further analysis could be performed to identify the reasons behind their playoff struggles. One possible limitation of this analysis is that it only focuses on the regular season performance and may not capture the unique challenges of playoff games, such as the pressure and intensity of a series. Moreover, future work could explore how the team can improve their playoff performance, such as through roster changes and coaching strategies. Overall, while the Dodgers' regular season performance was remarkable, there is always room for improvement and opportunities to learn from their playoff experience.