Heytea Everyday

Lineup Optimization:
Analytics on Lakers' Lineup Problem

ZENG Yun	**120090802**
GU Erxuan	**120020097**
SHEN Beiersha	**120020230**
LIU Xinyu	**120020128**
CHEN Ye	**120090035**
SHEN Hengyu	**120090633**
ZHONG Zhenyu	**120020317**

For the requirement of MGT4187

Dec. 13, 2022

1. Introduction

1.1 Background

Sports analytics is a heated application of machine learning and data mining. The major goal of sports analytics is to make an optimal decision and playing strategy to gain a competitive advantage. As one of the most popular sports in the world, basketball is a popular subject for sports analysis, and studies are mainly conducted in the NBA league. Most of the research on sports analysis topic has used machine learning methods to predict the outcome of the game and identify the significant factors that affect the outcome of games (Papageorgiou, 2022), and some research has found that the most critical indicators are: Win%, Offensive EFF, 3rd Quarter PPG, Win% CG, Avg Fauls, and Avg Steals. (Mikołajec et.,2013) In practical applications, these results connected with top teams and elite players are usually related to the lineup problem. All teams strive to bring their best performance to a game, which requires considering all the possible lineups. Therefore, determining the lineup is more and more significant for a team in their winning efforts. Spector (2020) constructed a machine learning model to predict actual lineup production in real games and mock a coach's decision-making process for creating a lineup based on individual statistics and how long a lineup will play. Since team performance is closely related to financial resources and other commercial income, more and more managers have attempted to use sports analytics to optimize the team's lineup. As many financial issues are involved, the analytics conducted on basketball team performance could be considered business analytics. Most studies have focused on the league as a whole, with little specific analysis of individual teams. Lineup issues could be very different from team to team.

As a legend in the NBA league, the Lakers have won 17 championships in its team history. In this regular season, season 2022-2023, the Lakers have astonishing performance in this season. With a winning rate of 35% (as of Nov. 27, 2022), the Lakers may not even enter the playoffs. With the Lakers in a slump, this project will use machine learning and data mining methods to optimize their lineup and improve their performance.

1.2 Data

The data used was scraped from NBA official stats, players and team statistics (basic and advanced), current rosters, and schedule results for all teams from the current 2021-2022 and 2022-2023 seasons. The data were all scraped and put into data frames. The data file form is 'csv' or 'xlsx' for the python applications. Player and team stats reflect all the recorded data of a player or a team on the field, respectively. The rows record the player's or team's name, while the columns correspond to various on-court performance data or identity data.

1.3 Methods Overview

Model one methodology uses data mining for NBA player and team analysis. For players, we use machine learning methods to fit player performance and player ability rating values from past seasons to predict the ability values of players so far this season. For the Lakers team, we use feature classification of data severity rating ranking to get the urgent team problems for the Lakers and give lineup replacement suggestions. Model two methodology uses the player recommendations derived from model one combined with decision trees and optimization methods for player lineup selection. Model 3 methodology uses stochastic simulation to simulate the lineups before and after the replacement for an entire season. It measures the accuracy of Model 2 by examining whether there is a significant performance improvement.

2. Model Construction and Results

2.1 Model I: Data Mining

The data mining has two parts; the first part is the player's ability analysis; we measure the players' ability value based on their performance so far this season through machine learning methods, which aim to observe whether there is an arbitrage opportunity (to get low salary and high ability players) in combination with salary. The second part is the problem analysis of the Lakers. We use different data classifications, visualization methods, and severity ranking to determine the urgent problems the Lakers need to solve. Through the data mining model, we were able to get a measure of the player's abilities and a refined analysis of the team's problems, which provided a theoretical basis for the model of player selection.

2.1.1 Part one

The idea of data mining for players analysis is to evaluate the players' ability based on their performance in last season. Based on the evaluation, we tend to find the arbitrage opportunity (to get low salary and high ability players) in combination with salary. We have the following steps to complete our analysis:

Data preparation

- Data collection
  - we have collected NBA players' performance information (traditional level, advanced level and misc level), biological information, salary information and NBA 2K scores in previous two seasons (2021-22 and 2022-23). All data is collected from the NBA.com which is the official database of NBA league.
- Data filtering
  - In order to avoid the effect of outliers (small number of games played with extremely good or bad performance), we have deleted the players' information which have attended less than 20 games in season 2021-22 to build a more robust model.
  - Min-max scaling has been done to avoid large weight difference in our model construction.
  - Result showing:

Table 1: Players' information in 2021-22

PLAYER	Team	Age	GP	W	Others	2K Scores
Joel Embiid	PHI	28	68	45	.....	95
LeBron James	LAL	37	56	25	.....	96
Kevin Durant	BKN	33	55	36	.....	96
Luka Doncic	DAL	23	65	44	.....	94

Table 2: Players' information in 2022-23

PLAYER	Team	Age	GP	W	Others	Scores
Stephen Curry	PHI	18	34.6	31.7	.....	?
Luka Doncic	LAL	16	37.1	34.0	.....	?
Kyrie Irving	BKN	12	36.0	25.1	.....	?
.....	.....	.....	.....	.....	.....	?

Data processing

- Model training and testing (regression choosing)

  Based on the players' information in 2021-22, we have put it into 10 regression methods to estimate players' scores and compared with the 2K scores to test the accuracy of the regression. With the training and testing scores:

Table 3: Regression method performance

Regression Method	Training Scores	Testing Scores
Linear Regression	0.846244	0.694946
Ridge method	0.817059	0.759697
Bagging Regressor	0.937380	0.749120
KNeighbors Regressor	0.810436	0.699067
SVR method	0.731888	0.581524
Lasso method	0.004152	-0.003536
MLP Regressor	0.862328	0.686144
Decision Tree Regressor	1.0	0.384726
Extra Tree Regressor	1.0	0.340967
Gradient Boosting Regressor	0.860429	0.792255

The bagging regression method has the relatively best performance in both training and testing part which gives the most accurate result in the regression choosing part.'

- Scores estimation

  Based on the chosen model from regression choosing part, we have put the players' information dataset of the 2022-23 season into the bagging regressor method and generated the estimated scores of players in the 2022-23 season.

Result analysis

- Estimated scores:

Table 4: Players' estimated scores in 2022-23

Rank	Player name	Scores
1	Luka Doncic	92.3
2	Joel Embiid	91.9
3	Giannis Antetokounmpo	91.8
4	Donovan Mitchell	91.3
5	Stephen Curry	91.3
6	Damian Lillard	90.7
7	Pascal Siakam	90.2
8	Jayson Tatum	90.1
9	James Harden	90
10	Shai Gilgeous-Alexander	90

- NBA official MVP Ladder (from NBA official):

Table 5: Nov 25th Kia MVP Ladder

Rank	Player name
1	Luka Doncic
2	Jason Tatum
3	Nikoal Jokic

Table 6: Nov 18th Kia MVP Ladder

Rank	Player name
1	Luka Doncic
2	Jason Tatum
3	Giannis Antetokounmpo

Table 7: Nov 11st Kia MVP Ladder

Rank	Player name
1	Giannis Antetokounmpo
2	Luka Doncic
3	Jason Tatum

Table 8: Nov 7th Kia MVP Ladder

Rank	Player name
1	Giannis Antetokounmpo
2	Luka Doncic
3	Donovan Mitchell

- Illustration:

  As the estimated rank has high similarity to the NBA official MVP ladder, the bagging regressor model is robust to estimate the scores of the players. Based on the estimated scores, we are more possible to find the arbitrage opportunity in combination with salary.

2.1.2 Part two

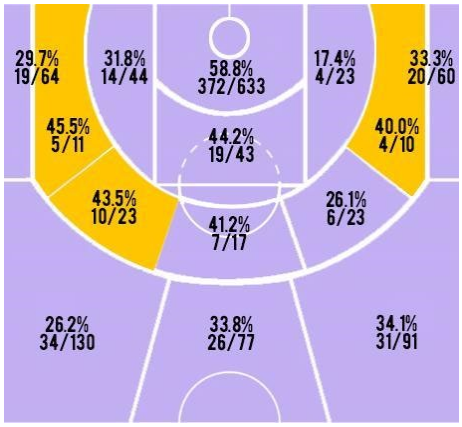
The idea of data mining for team analysis is to classify and rank different high-level data of teams according to team features like shooting or defense, etc. We observe the Lakers' ranking in 30 NBA teams under this feature classification; the ranking value of each feature class represents the severity. Finally, we propose the most crucial problems that the Lakers should solve first.

It is important to note that there are two types of recommendations made by data mining; one is to analyze the optimization strategies for lineup adjustments from the player's perspective. The other is to suggest optimizing potential tactics and other factors from the team's perspective. The recommendations from the lineup perspective in data mining will be used as the next step in the decision tree and optimization of the lineup restriction factor, while the analysis from the team perspective is called a possible impact factor because it has both subjective and objective considerations; it is

not considered for player selection, but still applies to the team's changes to the whole.

Microeconomics proposes to make decisions when always faced with trade-offs. The same is true for lineup changes; we cannot guarantee improved performance before and after lineup changes, so our data mining will only suggest the impact factors of different replacement scenarios. The effectiveness of the final decision is verified by the lineup optimization in Model II and the stochastic simulation in Model III.

- **Shooting Analysis:** We counted the Lakers' shooting rankings in each area of the court and then visualized the team's shooting performance; yellow areas mean above league average or equal, while purple areas mean below. Overall, the Lakers' offensive performance is abysmal, so they must find shooters who can shoot well.



- **Dribble Analysis:** Fewer dribbles (e.g., 0 or 1) mean quicker shots after the catch; players often have better shooting opportunities because the player's opposite defender may not be in a defensive position in time. But the Lakers were very poor in this area. This may explain why the Lakers' shooting was so bad in the last part of the game, as they didn't take good shots off the dribble instantly.

Table 9: Data Mining – Dribble Analysis of Season 2022-23

Dribble Times	ERG% Rank
0	28
1	25
2	22
3-6	15
7+	20

- **Tracking Pass Analysis:** The Lakers' assisted passing efficiency performs well, which shows that the Lakers have players with good passing ability, but the overall ball movement performs not well, probably because the team's tactics are not rich enough, so the Lakers need to enrich the tactical system and increase the ball movement.

Table 10: Tracking Pass Dashboard of Season 2022-23 and Season 2021-22

Dashboard Item	2022-23 Rank	2021-22 Rank	Rank Change
Passes Made	28	17	-11
Passes Received	28	17	-11
AST	18	17	-1
Secondary AST	24	23	-1
Potential AST	10	14	4
AST PTS Created	19	20	1
AST ADJ	17	19	2
AST to Pass%	6	17	11
AST to Pass% ADJ	4	17	13

- **Defensive Analysis:** The overall defensive performance is good, but there is a significant decrease in the team's defensive frequency and inside rebounding advantage compared to last season. So the Lakers team needs to improve their defensive aggressiveness and simultaneously enhance their inside rebounding ability.

Above all, from these characteristics analysis, we suggest three factors for player selection perspective: shooting percentage, 3-point shooting percentage, and offensive rebounding ability. Our recommendations for the overall team perspective are to enrich the play tactics system, enhance the ability to shoot quickly off the catch, and improve the team's defensive aggressiveness.

Table 11: Defense Conclusion of Season 2022-23 and Season 2021-22

Dashboard Item	2022-23 Rank	2021-22 Rank	Rank Change
DEF RTG	23	10	-13
DREB	3	12	9
DREB%	8	21	13
STL	18	12	-6
BLK	19	7	-12
OPP PTS OFF TOV	16	8	-8
OPP PTS 2nd Chance	12	12	0
OPP PTS FB	4	3	-1
OPP PTS Paint	11	3	-8
FREQ%	21	16	-5
DFGM	25	13	-12
DFGA	15	12	-3
DFG	27	17	-10
FG%	1	16	15
DIFF%	30	15	-15

2.2 Model II: Classification and decision tree

2.2.1 Indicators Classification

- **Model Construction Motivation:**

In order to search the specific players according to the ability indicators, the team focuses on. We need to construct an indicators classification model, which can help the team managers compare different players conveniently.

- **Model Description:**

This model sorts the value of each player's single indicator and divides the players into n groups. Every group contains 1/n * the number of all players (in practical operation, n equals 10). We define the best group as Class_10 and the worst group as Class_1. After sorting each indicator according to this method, we will get a list. This list will contain the league's overall ranking of a player's abilities. We can use this list to tell if the specific abilities of a player are good or bad and help team managers select players based on specific needs.

- **Model implementation:**

1. Indicator sorting function

   We can use Python to implement the sorting directly, and the sorting function code is in *M2-1-Indicators Classification.ipynb* (We have preprocessed the data. Some indicators are the lower, the better, such as the number of turnovers per game):

2. Indicators ergodic classification

   We use the sorting function for each piece of indicators. Then we can get the whole players' rank list (ALL_PLAYERS_Classification_Result.csv).

■ **Model Results:**

Finally, after the above implementation steps, we get the following rank list (Because of space is limited, we will use a few representative player examples):

Table 12: Indicator rank list

PLAYER	PTS	FG%	AST	REB	Others
LeBron James	10	9	10	10	.....
Luka Doncic	10	6	10	10	.....
Stephen Curry	10	5	8	10	.....

2.2.2 Team players screening

■ **Model construction motivation:**

In order to search the worst players through a team according to the ability indicators which the team focus on. We need to construct a screening model, which can help the team managers look for the traded players.

■ **Model description:**

This model requires us to take the output of Model II (weakness of a Team) as input, then the model will rate the players in team based on these indicators. Finally, the n players with the lowest score will be selected.

■ **Model implementation:**

1. Scoring function:

   We use Python to implement the scoring function and the code is in *M2-2-Team players screening.ipynb*.

   This scoring function will give a rank number of players based on the certain indicator, then the -rank_number will be his rank score.

2. Screening function:

   After sum all the rank scores, we can select the n worst players. Code is in *M2-*

2-Team players screening.ipynb.

■ **Model Results:**

In the case of the Lakers, we found these important indicators (['FG%','3P%','OREB']) in Model I (Data Mining). We follow the above method and set the n equal 3. Finally we get the 3 players with their ranking scores:

Table 13: The worst three players in the LAKERS ranking score

PLAYER	Score	FG%_score	3P%_score	OREB_score
Horton-Tucker	-33	-11	-13	-9
Kent Bazemore	-33	-14	-6	-13
Trevor Ariza	-37	-13	-12	-12

These three players will be selected to trade in next step: Player Arbitrage model.

2.2.3 Decision search tree - Player arbitrage

■ **Model construction motivation:**

This is the final and critical step of Model III, we will find the best substitute players with the help of Decision search tree. This model will optimize our trading payoff. At the same time, the model can well meet our various expected needs (Salary restrictions, minimum ability requirements and so on).

■ **Model description:**

The model needs to combine all the results we've had before, such as the indicator rank list, the traded players, the player ability overall score (Model I) .....

Then, we will use these information to construct the decision search tree, the tree's decision nodes can be adjusted, inserted or deleted. Every decision point can be explained by basketball lineup theory. In the end, the best substitute player for the Lakers at this stage will be selected by the model.

■ **Model implementation:**

After reviewing the previous steps, we now have the following constraints and requirements:

1) Salary constraint: The substitute player's salary cannot exceed the salary of the traded player.

2) Specific ability requirements: The substitute player's rank number of the

indicators we focus on must be higher than the traded player's.

3) Highest player's ability score condition: After the specific ability screening, if there is more than one player, we need to select the player with the highest ability score.

4) Special requirements of managers: We can easily meet the special requirements of the management by adding new nodes in the tree. Some special requirements example: The player's age, the player's nationality, the player, the previous team of the player and so on.

The Decision Search Tree code is in *M2-3-Decision search tree-Player arbitrage.ipynb*.

■ **Model Results:**

In the case of the Lakers, we construct the Decision search Tree and find the three best substitute players: 1. Jakob Poeltl; 2. Desmond Bane; 3. Isaiah Roby.

2.3 Model III: Stochastic simulation Testing

■ **Model construction motivation:**

After obtaining the ideal lineup for Lakers from previous models, the lineup is tested to see if the winning percentage was improved as expected and whether the performance has been significantly improved compared to the original lineup.

■ **Model description:**

The model is based on an open-source, stochastic simulator released by NBA officials. The data in this simulator are derived from real historical NBA data and player performance to simulate a full NBA season. This also includes random factors such as player injuries, contracts issues, trades, etc.

During the simulation, we allowed the CPU to perform coaching duties, including developing training plans, proposing or handling trades, and allowed for player injuries, contract expiry, etc., in the hope of achieving the maximum degree of random simulation.

Due to the operational limitations of the simulator, we can only obtain the season data after simulating 81 games (a real full season includes 82 games), but we do not think this affects how the data reflects the performance of the team.

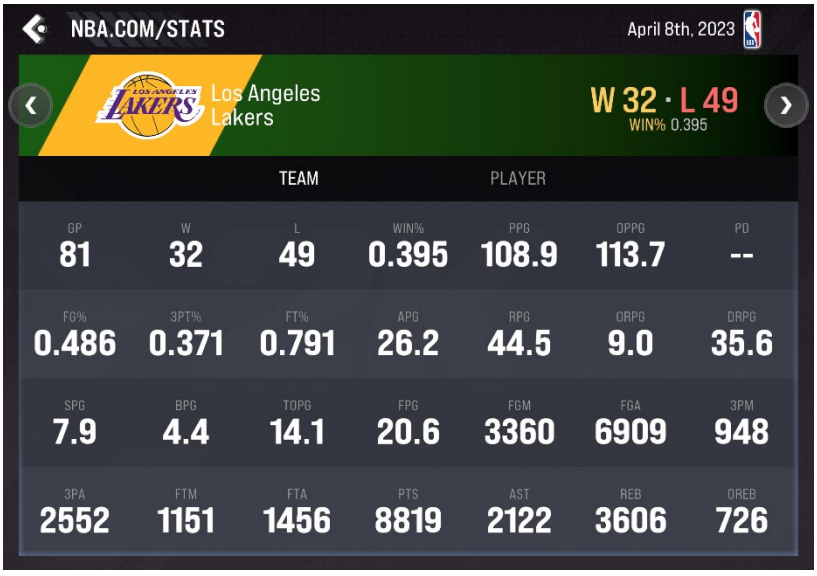
■ Model implementation:

1. Generate the controlled group of the original lineup

We put the original lineups into the simulator machine and run for 30 seasons. The original lineup is the one now used by Lakers in the real league, which consists of following players:

As mentioned in 2, we allowed the maximum number of random situations to occur, completing a simulation of 81 games over a season to obtain data on the team's performance.

One of the results is shown below:



Repeating the above process, we ran 30 simulations in total and obtained the following win rate results.

25.9	25.9	30.9	30.9	30.9	30.9	38.3	38.3	38.3	38.3
39.5	39.5	39.5	39.5	39.5	39.5	39.5	46.9	46.9	46.9
46.9	46.9	51.9	51.9	51.9	53.1	53.1	53.1	54.3	54.3

The mean value is **42.11%**

2. Generate the experimental group of the changed lineup

In the simulator, Lonnie Walker IV, Kendrick Nunn and Damian Jones left the Lakers in a trade that saw the Lakers get Isaiah Roby, Jakob Poeltl and Desmond Bane.

The changed lineup consists of following players:

We put the changed lineups into the simulator machine and run for another 30 seasons. Following results are what we obtained.

71.6	71.6	66.7	66.7	66.7	66.7	66.7	66.7	66.7	64.2
64.2	64.2	64.2	64.2	64.2	64.2	64.2	64.2	63	63
63	63	61.7	61.7	61.7	61.7	59.3	59.3	58	58

The mean value is **67.43%**

3. OLS Regression Results

After obtaining a total of 60 results of wining rate, we ran a regression analysis on this data set and obtained the following results:

■ **Hypothesis:**

The revised lineup does not significantly improve Lakers' performance.

	coef	std err	t-value	P>|t|	[0.025	0.975]
Intercept	48.3019	1.813	26.638	0	44.492	52.111
x_reg	0.3721	0.049	7.58	0	0.269	0.475

Omnibus	5.695	**Durbin-Watson**	1.166	**Skew**	1.088	**Prob (JB)**	0.136
Prob (Omnibus)	0.058	**Jarque-Bera**	3.99	**Kurtosis**	3.239	**Cond. No.**	220

- **T-test result:**

  t-statistic is: -10.2461; p-value is: 1.801^(-10)

  Therefore, we can reject the hypothesis.

- **Conclusion**:

  The winning rate has been significantly improved after the lineup change.

3. Limitation

The project focuses on evaluating players from a statistical perspective. Some authoritative measures reasonably combine players' information to calculate and compare their value. However, the project models have some limitations due to risks and uncertainty.

3.1 Player

- **Physical:** A player's physiology cannot be predicted. For example, players' injury risk cannot be predicted before or during a game and when they will go off injured or play because their teammate is injured.

- **Mental**

  - **Personal experiences and characters.** Personality is difficult to be quantified, and thus hard to predict the game results. Some players may be talented but are unwilling to take on supporting roles, leading to subpar team performance, such as the Warriors' Wiseman.

  - **Sentimental management ability.** Players' emotions are affected by various factors, and players' sentimental management ability is subjective. For example, cohesive rhetoric may motivate players, or make them too complacent about losing the game. The influence of the atmosphere in the locker room and people outside the arena on the players' emotions cannot be underestimated.

  - **Cooperation and tactical awareness:** Many cooperation behaviors are not statistically visible, such as "passing leads to assists" and "running to draw the

defense." However, measuring a player's ability and value is difficult because the relevant data is incomplete.

3.2 League and Coach

Different leagues have various training and playing styles. Coaches' status and tactics are dynamic on the field and hard to predict accurately. Consequently, the arrangement and performances of players vary in different situations.

3.3 Hardware Facility

Although regular games are strictly checked, facilities inevitably affect the game results. The variances in venues, facilities, equipment, and gear will cause game results that are inconsistent with the predictions.

3.4 Social Opinion

Social opinion may lead to team gaps between players within or between the teams. For example, Warriors' Green punched Pelicans' Poole, and the incident was not digested internally but amplified through social media, affecting both players' performances. Besides, players may experience the torment of public opinion. For players with high expectations accompanied by a certain degree of self-denial, social factors such as negative gossip and disparaging remarks can trigger unpredictable dejection or anger, affecting players' performances.

Overall, it is possible to predict and evaluate the performance of players and teams through quantitative analysis to some extent. However, we may be limited in comprehensively considering every factor affecting our predictions and decisions. Besides, some factors influencing players' value are difficult to quantifid under the current data analytics capabilities. The accuracy and effectiveness of predictions cannot be guaranteed only based on data and models for the uncertainty. However, this may be the charm of basketball and other competitive sports for the thrilling unknown results.

4. Conclusion

First of all, our data is based on the NBA official website, and the data processing

methods are as follows: Machine Learning, Data Mining, Classification, Decision Tree, and Stochastic simulation.

After analyzing replacing the players' lineup as the best improvement method for the NBA team, we build models to explore the optimal player lineup for the lakers.

First, the NBA players are classified in detail using Machine Learning, and their abilities are quantifiable under each classification and have a complete ranking. Then use scatter diagrams to organize players and find out the weak points of the lakers. The decision tree is used to find the most suitable high-quality replacement player when replacing players at the weak points of the Lakers. Then the Stochastic simulation machine (2K) was used to simulate the new players' lineup in the actual game many times, and we find the winning rate was greatly improved. The T-test is used to prove the robustness of the results. Finally, there is a hint about the limitations of the model.

So far, our main contributions are:

- Find the optimal player selection for the lakers. It can also flexibly use in every team.
- The black box of the neural network has been roughly explained to some extent.
- The ability of all NBA players has been classified and ranked in detail.

Suggestion

When applying the model to reality, some issues must be considered:

- The NBA draft is highly uncertain, and the team may not be able to obtain the idealist player in the transaction easily.
- The history of some teams also needs to be considered. The role of some players in the team will also become a symbol of the team without giving special consideration to their abilities.
- NBA is a zero-sum game on a competitive level. But on a business level, the unpredictability and watchability of the game, the fun of the draft, are the main points that can benefit the league, and it's a win-win game. So, in reality, what this model can give is idealistic advice. The rest depends on fate and luck.

References

Mikołajec, K., Maszczyk, A., & Zając, T. (2013). Game indicators determining sports performance in the NBA. *Journal of human kinetics*, *37*, 145.

Papageorgiou, G. (2022). Data Mining in Sports: Daily NBA Player Performance Prediction.

Spector, J. (2020). *Optimizing NBA Lineups* (Order No. 27999920). Available from ProQuest Dissertations & Theses Global; ProQuest Dissertations & Theses Global A&I: The Humanities and Social Sciences Collection; ProQuest Dissertations & Theses Global A&I: The Sciences and Engineering Collection. (2421174029). https://www.proquest.com/dissertations-theses/optimizing-nba-lineups/docview/2421174029/se-2