CONCEPTUAL PAPER

Arun kumar Maram

ADTA-5940 CAPSTONE PROJECT

Denise R. Philpot

Fall 2023

Table Of Contents

Chapter : 1 Introduction

Basketball is a vibrant, fast-paced sport that draws spectators from all around the world with its elite teams and exciting contests. The National Basketball Association (NBA) is the highest level of professional basketball, with teams competing fiercely and exhibiting extraordinary athleticism and strategic prowess. Our research aims to utilize important variables including team Elo ratings, past performance information, and the possible impact of rival team Elo rankings in order to comprehend and forecast NBA game results.

Our study's main goal is to create a reliable prediction model for NBA game results. This study is motivated by the premise that our ability to anticipate game outcomes will be significantly improved by incorporating team Elo rankings, historical performance data, and opposing Elo rankings into our predictive model. This theory emphasizes how important these factors are as trustworthy markers for deciphering the intricate patterns of NBA match results.

Beyond the domain of forecasts, our study also seeks to clarify the idea of an NBA "away game disadvantage." This intriguing theory proposes that when teams play away from home, their performance could suffer. We set out to investigate team Elo ratings with a focus on away games in order to find trends that indicate whether teams perform better or worse on a regular basis in these situations. Finding statistical proof of a "Away Game Disadvantage" and identifying differences in team performance in various game environments are the main objectives.

Context and Background

We use a study from the 2019 Men's Basketball World Cup to contextualize our research. This study looked at traits that elite teams had in common, studied attacking and defensive performance differentials, and connected these factors with the competition's final

standings. This interesting study used a variety of statistical techniques, such as correlation analyses and the rank-sum ratio (RSR), to explore the nuances of team performance in the international basketball arena.

Research Questions

According to our premise, the key to deciphering a deeper understanding of NBA game outcomes lies in the incorporation of team Elo ratings, historical performance indicators, and opposing Elo rankings into our prediction model. In addition, our investigation of a "away game disadvantage" seeks to provide insightful analyses of team dynamics and a nuanced viewpoint on the difficulties teams encounter in various playing conditions.
1. To build an accurate prediction for game outcomes (win/loss) using team Elo ratings, previous performance, and perhaps opposition Elo rankings?

2. To identify evidence of an 'away game disadvantage' in the NBA, can we analyze the progression of team Elo ratings in away games and identify teams who consistently perform better or worse when playing away from their home court?

In the sections that follow, we will methodically break down our theories, outline our approaches, reveal the findings of our investigations, and have meaningful conversations in order to draw important conclusions. In addition to improving our ability to forecast NBA games, the future path aims to provide new knowledge that will benefit basketball players, scholars, and strategists alike.

Chapter : 2 Literature

All ball team sports have similar goals, and basketball is no exception. Achieving sustained high-level performance is a challenging endeavour. The pressures placed on athletes during competitive seasons have been examined using earlier research approaches, such as match and time-motion studies (Sun et al., 2022). Interestingly, findings from research on basketball match analysis show that winning teams typically perform better than losing teams when it comes to making field goals and grabbing defensive rebounds. These results highlight how crucial field goal ability and strategic decision-making are to a team's success.

According to an interesting theory that was first presented in a study examining 870 regular season basketball games from 2000 to 2006, basketball season winners perform better defensively and in passing than game winners, who prioritize shooting accuracy. Ball steals and blocked field goals were two defensive metrics that were examined; these were activities that depended on the players' assertiveness and level of fitness (Sun et al., 2022). Although these studies provide important light on the dynamics of winning teams' performances, there is still a dearth of knowledge regarding seasonal variations in game-related data, especially when it comes to fitness factors.

Fitness factors: Researchers found that soccer players' physical performance and fitness levels varied significantly during the season, with notable gains during the latter part of the season. However, the majority of the research on seasonal variation in basketball that is currently accessible focuses on fitness testing and reports on generally insignificant or modest variations that occur both within and between competitive seasons. So, the question is: Could variations in fitness affect the technical performance of games?

The authors' response to this query emphasized that although overall fitness changes were insignificant, it is important to consider any possible consequences on game technical

performance. A player that experiences less fatigue during the season, for example, can demonstrate enhanced rebounding numbers, defensive preparedness, decision-making, passing, and shooting abilities.

Seasonal variations in game-related data in basketball are still largely unknown, despite the fact that replacement regulations differ between basketball and soccer. Using mixed modelling, researchers estimated mean changes in basketball competitive seasons as well as across them, concentrating on fitness variables like power and aerobic fitness. The conclusion that mean fitness within and between seasons shows minimal overall change was reached based on the results, which revealed that fitness changes were typically minor or tiny.

Although athletes frequently sense accumulated weariness, there is contrasting information in the literature on fitness testing that calls into question this idea. The fitness levels in high-performance basketball programs do not significantly alter during the competitive season, according to research. There are interesting considerations regarding the effect of felt exhaustion on technical abilities and choice reaction time that are raised by the discrepancy between athletes' perceptions and objective fitness assessments.

The study also explores the connection between different fitness metrics and playing time over a basketball season. Throughout the season, non-starters underwent aerobic detraining in contrast to starters, which could have resulted in increased weariness and a larger chance of defensive fouls. This result emphasizes how playing time, fitness levels, and in-game performance interact in a complex way.

Moreover, the examination of variations in game-related metrics between players who started and those who did not indicated that defensive actions, such as fouls committed and defensive rebounds, differentiated between both groups of players. This disparity begs the question of how well-conditioned the players are; it suggests that the starters may be more fit,

which would affect their ability to jump, defend and grab rebounds, and their aerobic capacity (Sampaio et al., 2010).

Building on this foundation, the current study seeks to determine within-season variations in basketball players' game-related statistics according to playing time and team quality. This study aims to provide new insights into the complex relationship between playing time, team quality, and seasonal fluctuation in basketball statistics by utilizing a variety of data sources. Examining how performance profiles vary based on playing time and team quality may help coaches and players get a deeper understanding of the basketball game. For instance, an intermediate team has to improve in passing and 2-point field goals, while a poorer team needs to improve primarily in defensive rebounding. A less significant player might also gain from concentrating on making fewer mistakes (Sampaio et al., 2010). The results of this study add to the body of knowledge already available on basketball performance, with a particular emphasis on how game location—home or away—affects game-related statistics and the results of professional men's basketball games played in the ACB League. Previous studies have examined the role that game-related statistics have in differentiating between winning and losing teams. Successful 2-point field goals and defensive rebounds are frequently mentioned as examples of these differentiating elements (“Study of Game-Related Statistics Which Discriminate Between Winning and Losing Basketball Junior Teams U-17 in World Championship,” 2014).

The results of this study support the widely held belief that defensive rebounds are essential to a team's performance. Defensive rebounds help the winning team perform better overall by limiting opponents' chances to score and allowing them to start productive offensive plays. Furthermore, the focus on assists as a differentiator lends credence to the notion that effective basketball play is mostly dependent on cohesive collaboration and strategic decision-

making. Sports research has shown interest in the home-advantage effect, which is the effect of team location on performance. The current study sheds more light on this phenomena by emphasizing the particular game-related metrics that distinguish winning and losing teams both at home and away. Previous research has indicated that a variety of factors, including familiarity, travel, and audience support, can affect how well a team performs in a given setting (Gómez et al., 2008).

Notably, the study's conclusions show that, depending on the game's location, the importance of specific game-related information fluctuates. Assists, for example, and made 2-point and 3-point field goals were found to be critical components of winning teams on the road. This implies that in order to win in strange settings, traveling teams must implement unique tactics like aggressive defense and effective shot selection.

The home court effect has been studied in the past, along with its relationships to factors like fouls committed and made free throws. However, this study offers a more nuanced understanding of the role defensive rebounds and assists play in home wins. The authors contend that home teams' aggressive play and the psychological benefit of playing in familiar settings are factors in their success (Gómez et al., 2008).

A study states, there is an increasing amount of research that highlights the value of game-related statistics in analysing basketball performance, especially when it comes to differentiating between winning and losing teams according on the location of the game. In order to fully comprehend the home court effect and its consequences for coaching and training strategies, more research across leagues and seasons is required(Madarame, 2018).

Chapter : 3 Methods

Sample and Data Collection: We used historical NBA game records from the 1946–1947 season, which was a subset of the dataset provided. The dataset contains statistics on game results, elo ratings, and team performance. Since the information was gathered from official NBA records, our analysis was accurate and trustworthy.

Dataset Description:

The dataset provides a thorough view for exploratory data analysis with 23 columns that contain specific information on NBA games. NBA games' play orders are shown in the "gameorder" column, and each game's unique identifier is shown in the "game_id" column. The league in which the game was played is indicated by the "lg_id" column, while the "_iscopy" column indicates if the opposing team has played the same game in the same matchup. "date_game" is the game date, and "year_id" is the season ID, which is called after the year the season ended. While "team_id" and "fran_id" offer three-letter codes and franchise IDs, respectively, for the team names, the "is_playoffs" column acts as a marker for playoff games. Other noteworthy columns include "elo_i" and "elo_n" for the team's elo before and after the game, "pts" for the team's point total, and "win_equiv" for the equivalent amount of victories for a team of elo_n quality in an 82-game season. "game_location" indicates whether the game was played at home (H), away (A), or in a neutral venue (N). "opp_id" and "opp_fran" stand for the opponent's team and franchise IDs. The "team_id" column's "game_result" column shows the team's result in the match, while the "forecast" column gives the Elo-based probability of winning based on the match's location and elo ratings. A "notes" column in the dataset contains further information. This extensive dataset provides a multitude of data for performing in-depth exploratory data analysis on the results and statistics of NBA games.

Figure 1

Dataset Used

	gameorder	game_id	lg_id	_iscopy	year_id	date_game	seasongame	is_playoffs	team_id	fran_id	...	win_equiv	opp_id	opp_fran	opp_pts
0	1	194611010TRH	NBA	0	1947	11/1/1946	1	0	TRH	Huskies	...	40.294830	NYK	Knicks	68
1	1	194611010TRH	NBA	1	1947	11/1/1946	1	0	NYK	Knicks	...	41.705170	TRH	Huskies	66
2	2	194611020CHS	NBA	0	1947	11/2/1946	1	0	CHS	Stags	...	42.012257	NYK	Knicks	47
3	2	194611020CHS	NBA	1	1947	11/2/1946	2	0	NYK	Knicks	...	40.692783	CHS	Stags	63
4	3	194611020DTF	NBA	0	1947	11/2/1946	1	0	DTF	Falcons	...	38.864048	WSC	Capitols	50
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	..
126309	63155	201506110CLE	NBA	0	2015	6/11/2015	100	1	CLE	Cavaliers	...	60.309792	GSW	Warriors	103
126310	63156	201506140GSW	NBA	0	2015	6/14/2015	102	1	GSW	Warriors	...	68.013329	CLE	Cavaliers	91
126311	63156	201506140GSW	NBA	1	2015	6/14/2015	101	1	CLE	Cavaliers	...	60.010067	GSW	Warriors	104
126312	63157	201506170CLE	NBA	0	2015	6/16/2015	102	1	CLE	Cavaliers	...	59.290245	GSW	Warriors	105
126313	63157	201506170CLE	NBA	1	2015	6/16/2015	103	1	GSW	Warriors	...	68.519516	CLE	Cavaliers	97

126314 rows × 23 columns

Figure 2

Dataset Used

There are 126314 rows and 23 columns in our dataset

	gameorder	_iscopy	year_id	seasongame	is_playoffs	pts	elo_i	elo_n	win_equiv	opp_p
count	126314.000000	126314.000000	126314.000000	126314.000000	126314.000000	126314.000000	126314.000000	126314.000000	126314.000000	126314.0000
mean	31579.000000	0.500000	1988.200374	43.533733	0.063857	102.729982	1495.236055	1495.236055	41.707889	102.7299
std	18231.927643	0.500002	17.582309	25.375178	0.244499	14.814845	112.139945	112.461687	10.627332	14.8148
min	1.000000	0.000000	1947.000000	1.000000	0.000000	0.000000	1091.644500	1085.774400	10.152501	0.0000
25%	15790.000000	0.000000	1975.000000	22.000000	0.000000	93.000000	1417.237975	1416.994900	34.103035	93.0000
50%	31579.000000	0.500000	1990.000000	43.000000	0.000000	103.000000	1500.945550	1500.954400	42.113357	103.0000
75%	47368.000000	1.000000	2003.000000	65.000000	0.000000	112.000000	1576.060000	1576.291625	49.635328	112.0000
max	63157.000000	1.000000	2015.000000	108.000000	1.000000	186.000000	1853.104500	1853.104500	71.112038	186.0000

Figure 3

Dataset Used

win_equiv	opp_pts	opp_elo_i	opp_elo_n	forecast
126314.000000	126314.000000	126314.000000	126314.000000	126314.000000
41.707889	102.729982	1495.236055	1495.236055	0.500000
10.627332	14.814845	112.139945	112.461687	0.215252
10.152501	0.000000	1091.644500	1085.774400	0.020447
34.103035	93.000000	1417.237975	1416.994900	0.327989
42.113357	103.000000	1500.945550	1500.954400	0.500000
49.635328	112.000000	1576.060000	1576.291625	0.672011
71.112038	186.000000	1853.104500	1853.104500	0.979553

Missing Values

Observations across the 23 columns are full and contain data for all specified variables, as evidenced by the dataset's lack of missing values. This completeness is essential to guaranteeing the validity of any modeling or analysis performed on the dataset, since missing values have the potential to introduce bias and jeopardize the correctness of findings. When a dataset is comprehensive, analysts and researchers can proceed with confidence when examining relationships, patterns, and trends since they are certain that no informational gaps exist that might compromise the validity of their conclusions. The complete integrity of the dataset is enhanced by the lack of missing values, which also makes exploratory data analysis more reliable and accurate.

Figure 4

Missing Values in the Dataset

```
gameorder             0
game_id               0
lg_id                 0
_iscopy               0
year_id               0
date_game             0
seasongame            0
is_playoffs           0
team_id               0
fran_id               0
pts                   0
elo_i                 0
elo_n                 0
win_equiv             0
opp_id                0
opp_fran              0
opp_pts               0
opp_elo_i             0
opp_elo_n             0
game_location         0
game_result           0
forecast              0
notes            120890
dtype: int64
```

Dependent and Independent variables

The dependent variable in the study of NBA game outcomes is "game_result," a binary indicator that shows if a team won (W) or lost (L). This variable is the main focus for evaluating the predictive models' accuracy. The approach takes into account independent variables such as team Elo ratings, which indicate the relative strength of home and away teams based on past performance. To adjust for current form, other criteria are included, such as points scored (pts), which indicate the team's recent performance. In order to provide an understanding of the level of competition encountered, the Elo ratings of the opposing teams are also taken into account. In order to investigate the effects of playing at home or away on game outcomes, a binary variable representing the game location (H for home, A for away) is also included. When combined, these variables provide a complete set of predictors for the analysis of NBA game outcomes, enabling a detailed look at the determinants affecting a team's success or failure.

Below is a distribution of game variables according to year. Variations in the total number of games played over time can be seen in the distribution of games between various years. According to the data, the years 2014 and 1972 had much more games than the other years in the dataset. This data points to possible changes in the NBA seasons' schedule or format in those particular years. The historical dynamics of the NBA may be better understood by examining the factors—such as expansion teams, modifications to the playoff structure, or other league-related events—that contributed to the increased number of games in these years.

Figure 5

Number of Teams played in each year

Dataset variables Distribution:

The following graphic aids provide in-depth explanations of the distributions of all the variables in the dataset. By giving us a more complex grasp of the patterns and traits of each variable, these plots are essential in enhancing our understanding of the individual variables. Through close examination of these representations, we hope to obtain a complete picture of the data and make better decisions as we use analytical techniques to investigate and test the hypotheses. These visualizations are an invaluable resource for revealing the underlying patterns and trends observed in the dataset. They provide us with the knowledge we need to make informed decisions as we explore and analyze the developed hypotheses.

Figure 6

Distribution of all the Data columns

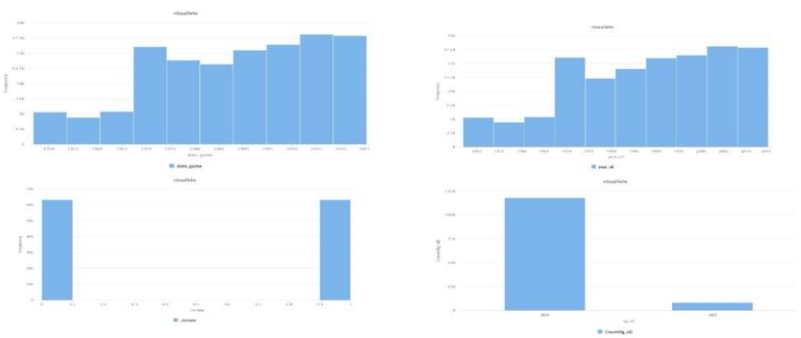
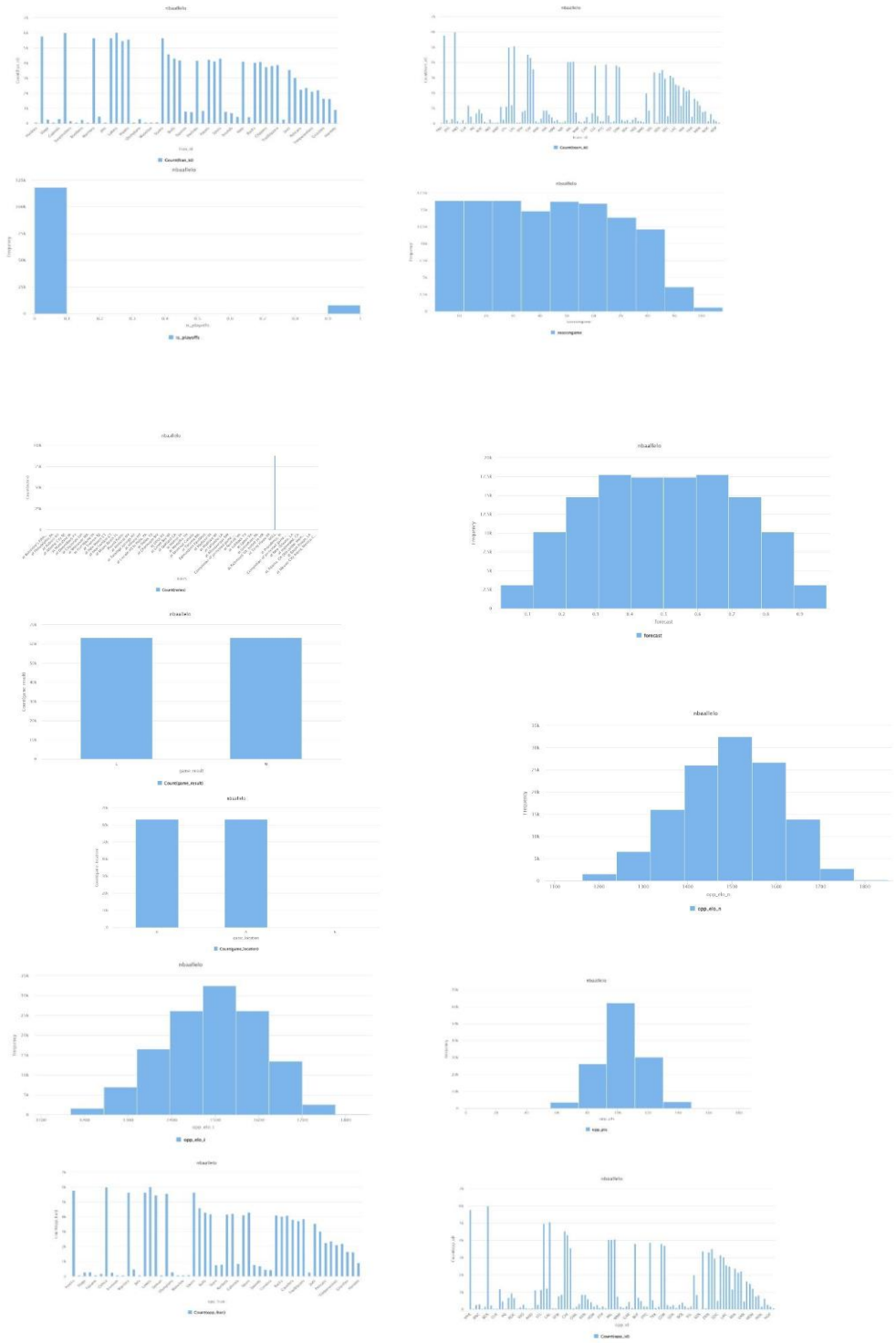


Figure 6

Distribution of all the Data columns



Correlation analysis of the dataset

Understanding the connections between the dataset's numerous numerical columns is possible thanks to the correlation matrix (Kubayi & Larkin, 2022). Upon examining the data, we find that the 'year_id' and 'gameorder' have a strongly positive correlation (about 0.99), suggesting a strong linear association between the season year and the play order of NBA games throughout history. A moderately positive connection has been seen between the 'is_playoffs' column and 'elo_i', 'elo_n', 'win_equiv', and 'forecast'. This indicates that playoff games are linked to greater team elo ratings, win equivalent values, and predicted winning chances. 'Win_equiv' and 'points_scored' show a noteworthy positive correlation, indicating that teams with greater point totals also typically have more victories in an 82-game season. The correlation between opponent points ('opp_pts') and win equivalent ('win_equiv') is negative, on the other hand, suggesting that teams with higher win equivalents typically give up less points. The 'elo_i', 'elo_n', and 'win_equiv' columns positively correlate with the 'forecast' column, which indicates Elo-based possibilities of victory; this suggests that better team elo ratings and win equivalents translate into higher anticipated winning probabilities. In general, the correlation matrix offers insightful information about the connections within the dataset.

Table 1

Correlation Matrix

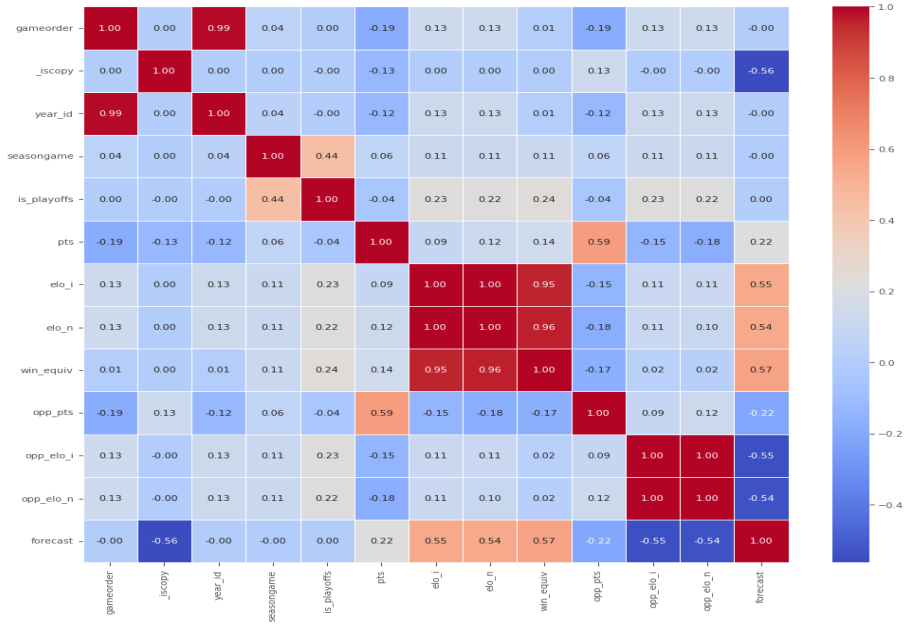
	elo_n	win_equiv	opp_pts	opp_elo_i	opp_elo_n	forecast
gameorder	0.130139	0.011238	- 0.186835	0.130512	0.130139	-3.396295e-11
year_id	0.125970	0.009217	- 0.123959	0.126331	0.125970	-2.615831e-11
is_playoffs	0.224773	0.236597	-0.039438	0.225418	0.224773	1.412979e-11
pts	0.121670	0.139801	0.592491	-0.147279	-0.178553	2.178615e-01
elo_i	0.996053	0.953463	-0.147279	0.105460	0.106245	5.460675e-01
elo_n	1.000000	0.957831	-0.178553	0.106245	0.099144	5.432202e-01
win_equiv	0.957831	1.000000	-0.173759	0.024532	0.017365	5.679823e-01
opp_pts	- 0.178553	- 0.173759	1.000000	0.090233	0.121670	-2.178615e-01

opp_elo_i	0.106245	0.024532	0.090233	1.000000	0.996053	-5.460675e-01
opp_elo_n	0.099144	0.017365	0.121670	0.996053	1.000000	-5.432202e-01
forecast	0.543220	0.567982	-0.217861	-0.546068	-0.543220	1.000000e+00

Table 2

Correlation Matrix

	gameorder	year_id	is_playoffs	pts	elo_i \
gameorder	1.000000e+00	9.889661e-01	3.666710e-03	-0.186835	0.130512
year_id	9.889661e-01	1.000000e+00	-3.853061e-03	-0.123959	0.126331
is_playoffs	3.666710e-03	-3.853061e-03	1.000000e+00	-0.039438	0.225418
pts	1.868347e-01	-1.239592e-01	-3.943777e-02	1.000000	0.090233
elo_i	1.305120e-01	1.263310e-01	2.254180e-01	0.090233	1.000000
elo_n	1.301386e-01	1.259696e-01	2.247731e-01	0.121670	0.996053
win_equiv	1.123794e-02	9.216879e-03	2.365967e-01	0.139801	0.953463
opp_pts	-1.868347e-01	-1.239592e-01	-3.943777e-02	0.592491	-0.147279
opp_elo_i	1.305120e-01	1.263310e-01	2.254180e-01	-0.147279	0.105460
opp_elo_n	1.301386e-01	1.259696e-01	2.247731e-01	-0.178553	0.106245
forecast	-3.396295e-11	-2.615831e-11	1.412979e-11	0.217861	0.546068

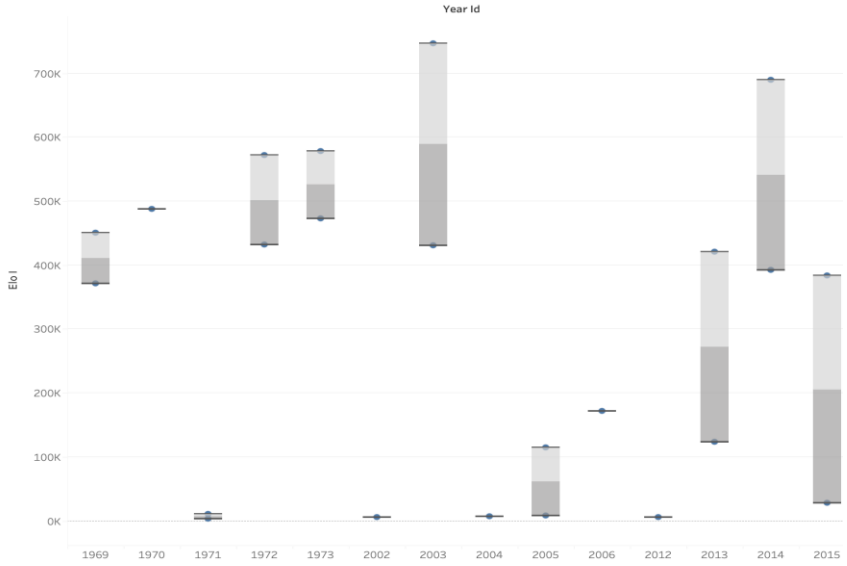


Weather Elo ratings effects the outcome of the game?

In 2003 and 2014, the NBA teams had greater Elo_i ratings as they started the game, indicating a better track record. These better ratings can be attributed to a variety of factors, such as skillful individuals, competitive league changes, team makeup changes, or dominant team performances. Contextualization and additional study are required to determine the causes of these higher evaluations.

Figure 7

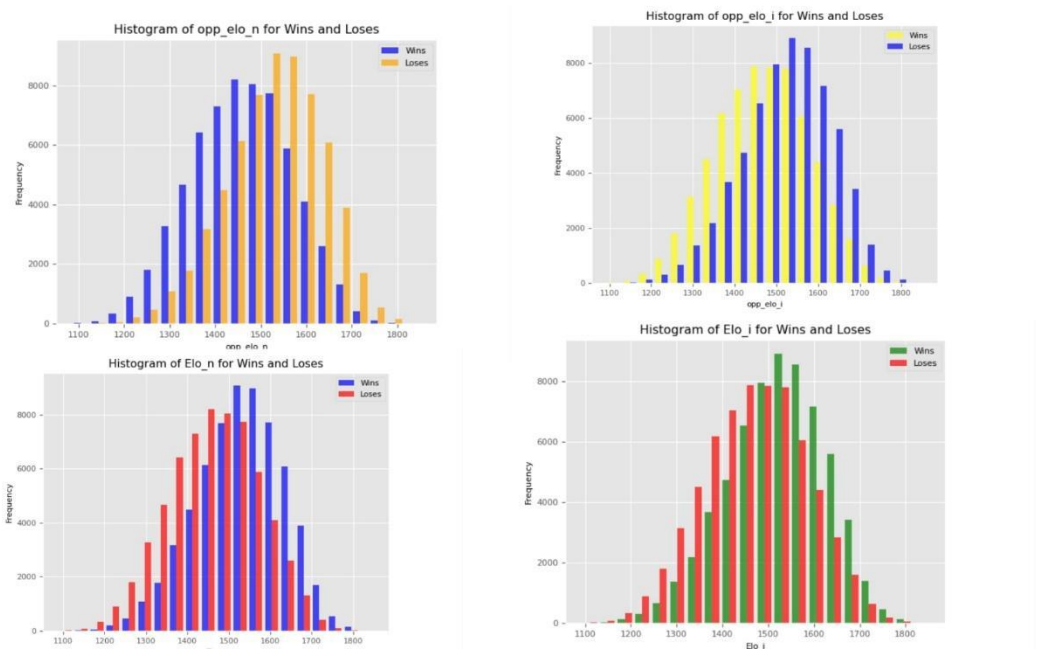
Distribution of Elo_i rating verses Year



In terms of wins and losses, the histogram distributions reveal a correlation between Elo_i, Elo_n, opp_elo_i, and opp_elo_n. There is a normal distribution. While teams with lower ratings may suffer more defeats, teams with higher Elo ratings typically win more games (Madarame, 2017). This implies that game results may be impacted by the opponent's strength, as indicated by their Elo ratings. This emphasizes how crucial it is to comprehend how Elo ratings and game results relate to one another.

Figure 10

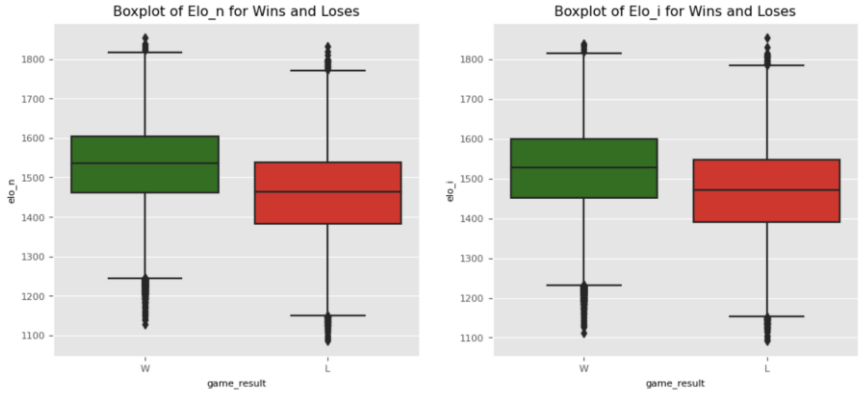
Distribution of elo_ratings verses Wins and loses



Elo_n and Elo_i boxplots compared to game_results demonstrate comparable distributions for various outcomes. This implies that the distributions of the original Elo rating and the revised Elo rating for winning and losing outcomes are comparable. This suggests that, independent of the outcome of a game, changes in Elo rating occur according to a same pattern. Boxplots can be used to evaluate outliers, the distribution of Elo ratings, and the general consistency of game results. Deeper insights into the evolution of Elo ratings and characteristics impacting team success might be obtained by doing more research or taking into account other variables.

Figure 11

Boxplot of Elo_n and Elo_i ratings vs game_result



The dataset's distribution 'points' feature and elo_i, which is a relatively normal distribution between values 0 and 200 points, is displayed in the graph below.

Figure 12

Distribution of points column

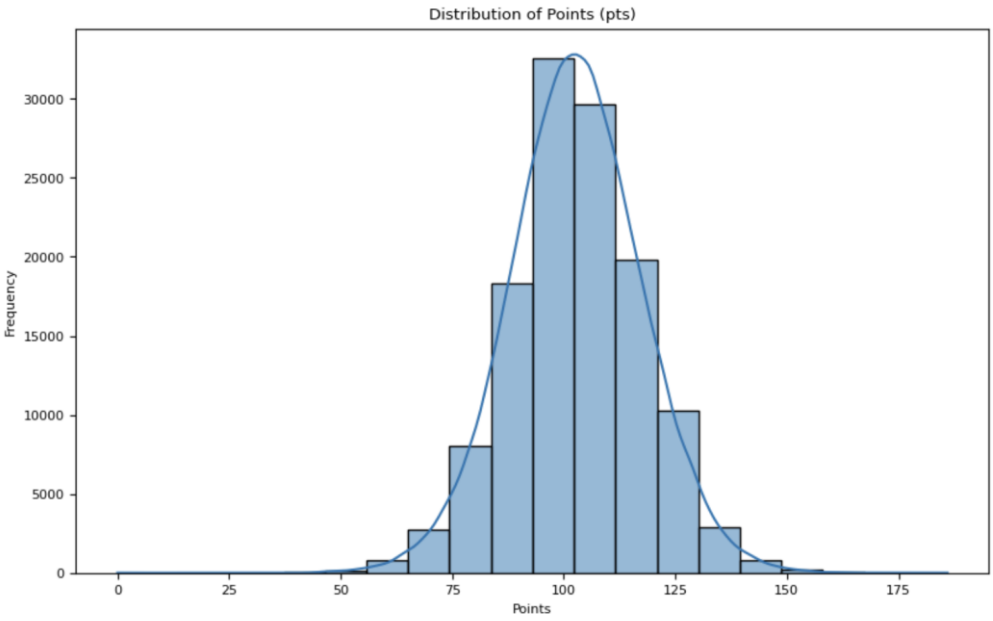
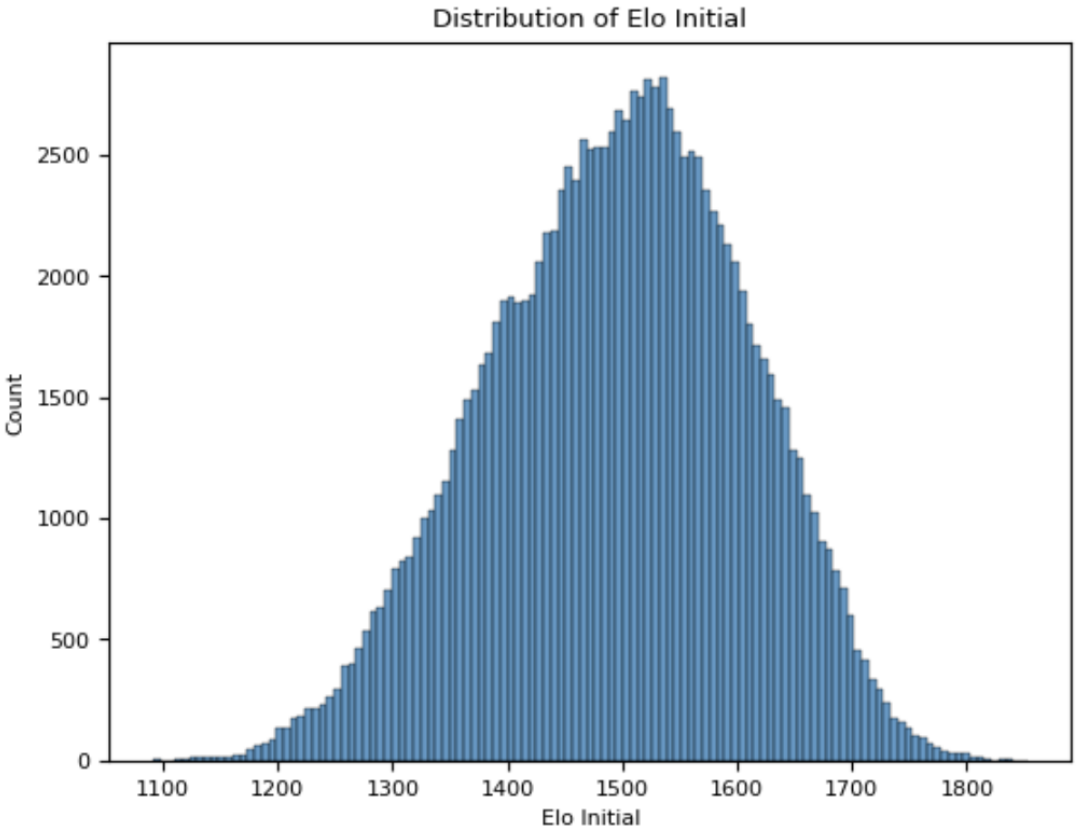


Figure 13

Distribution of points column



Random Forest Method

To fulfil the first research goal, determine whether the game results are influenced by prior performance or elo ratings. We have developed a random forest forecasting model that uses the opposition's Elo rankings, team Elo ratings, and recent performance as predictors. With an accuracy of 98.25%, the random forest model's remarkable performance metrics highlight how well the selected features—such as Elo ratings and past performance data—predict game outcomes. The model is able to distinguish between winning (1) and losing (0) outcomes with greater effectiveness, as demonstrated by the precision, recall, and F1-score measures. A balanced and accurate prediction for both winning and losing scenarios is indicated by the excellent precision and recall ratings for both classes (Madarame, 2017). The hypothesis that

Elo ratings and past performance have a major impact on game results in the NBA dataset is highly supported by these results. The remarkable accuracy of the model implies that the selected variables identify significant trends in the data, demonstrating its applicability in forecasting basketball games. This result gives stakeholders important information by showing that Elo ratings and past team performance are trustworthy predictors of game outcomes.

Figure 13

Performance report random forest model

Accuracy: 0.9824596135571745
Classification Report:

	precision	recall	f1-score	support
0	0.98	0.98	0.98	12691
1	0.98	0.98	0.98	12565
accuracy			0.98	25256
macro avg	0.98	0.98	0.98	25256
weighted avg	0.98	0.98	0.98	25256

The second research issue, "away game disadvantage," will be examined using a different methodology (Madarame, 2018). We'll examine how a team's Elo ratings fluctuate between away and home games. The degree to which a team routinely performs well or poorly away from home will determine their ranking.

A hypothesis test is suggested to investigate the link between the game outcome (win or loss) and the difference in Elo ratings (elo_n - elo_i) in the context of the dataset that is provided. The null hypothesis (H0) states that there is no meaningful correlation between changes in Elo ratings and the likelihood of winning or losing a game, indicating that variations in Elo ratings have no effect on the odds of winning or losing. The alternative hypothesis (H1), on the other hand, contends that there is a substantial correlation between the variation in Elo

ratings and game outcomes, suggesting that shifts in Elo ratings affect the likelihood of winning or losing. The alternative hypothesis contends that at least one of these probability is different from the null hypothesis, which holds that the conditional probabilities of winning or losing given the variation in Elo ratings are identical. In order to shed light on the significance of Elo ratings in forecasting game results within the context of the dataset, this hypothesis test attempts to give statistical evidence about the impact of Elo rating changes on game outcomes.

Will the game outcome depends on the Game location?

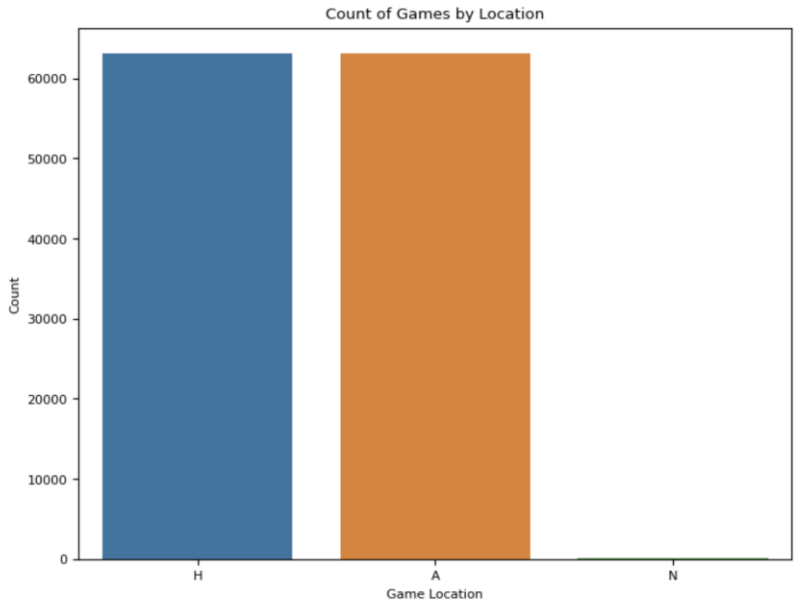
The 'away game disadvantage' will be the subject of an analysis to answer the second research question. Examining the development of a team's Elo ratings, particularly when playing away from home, is part of this. Our goal is to find patterns of constant dominance or inferiority for teams playing away from home by analyzing how team performance, as measured by Elo ratings, changes in various game environments. The study question suggests that teams may typically be at a disadvantage while playing away from home. The investigation will clarify this. This study of the dynamics of team performance in various game settings offers insightful information about how the playing environment affects team performance and offers a detailed knowledge of the potential difficulties that come with playing away from home.

The number of games played at home and away seems to be equal in the dataset, with roughly 60,000 games played in total at each venue. The dataset appears to be evenly distributed with regard to game locations based on the ratio of games played in both home and away environments. This equal distribution ensures that any observed differences in performance are less likely to be influenced by a substantial gap in sample size and offers a solid foundation for assessing and comparing results across home and away games. In order to reduce the influence of confounding variables and improve the validity of statistical studies,

researchers frequently strive for balanced datasets. In this instance, the evenly split distribution of home and away games makes it easier to determine how much the game's location influences the quantity of game.

Figure 14

Game location distribution



A bar graph was made to show the difference in wins between home and away games, and it showed a clear pattern of home games having more wins than away games. The two categories, "Home" and "Away," on the graph's horizontal axis, stand for the corresponding game locations. The quantity or frequency of victories in each category is shown on the vertical axis. There is a noticeable difference in the height of the bars labeled "Home" and "Away," suggesting that more games were won at home. This graphical representation offers a clear visual understanding of how home teams outperform away teams when it comes to winning matches. The difference in the bar heights highlights how important home-court advantage is in the dataset and implies that teams usually have better results while playing at home.

Figure 15

Game results with respect to Game location

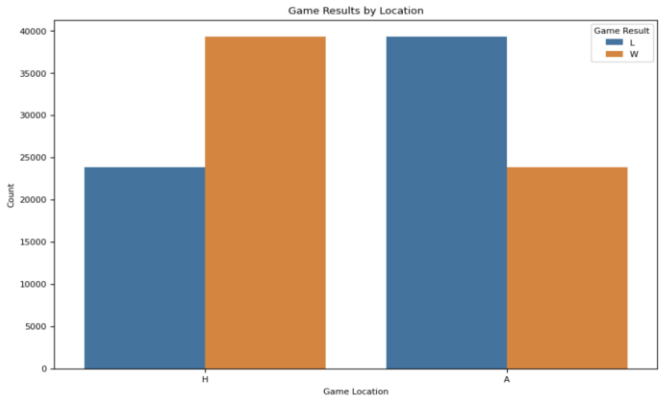


Figure 16

Distribution of numerical columns

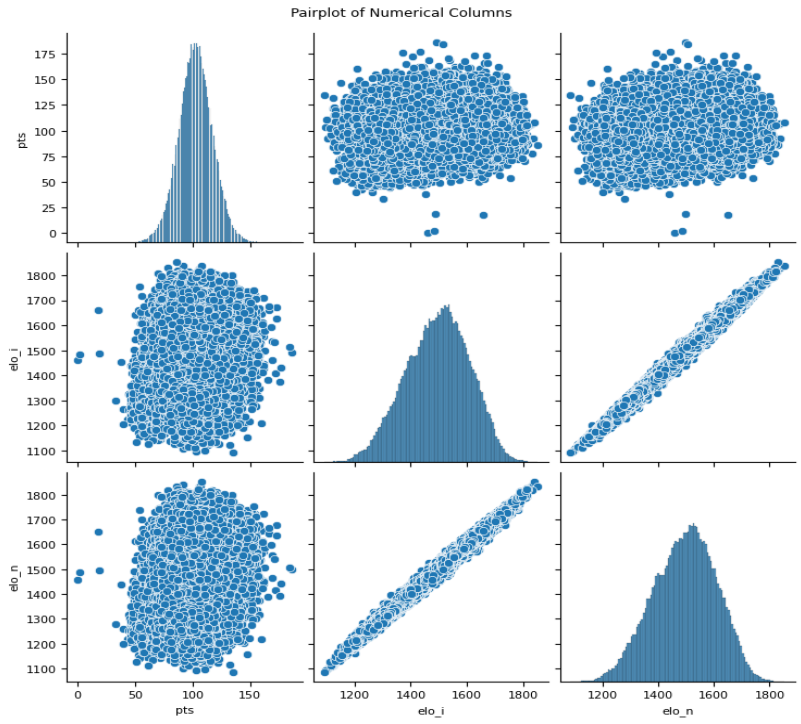


Figure 17

Game result distribution with respect to year

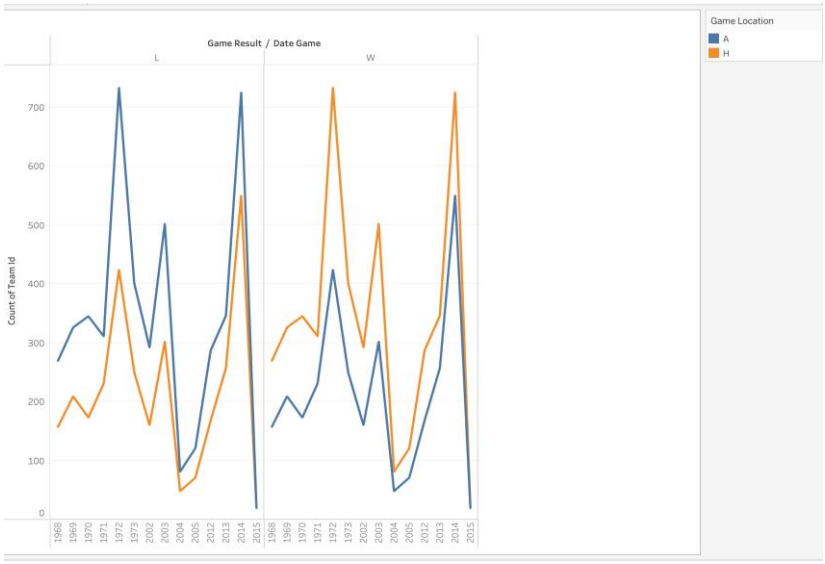
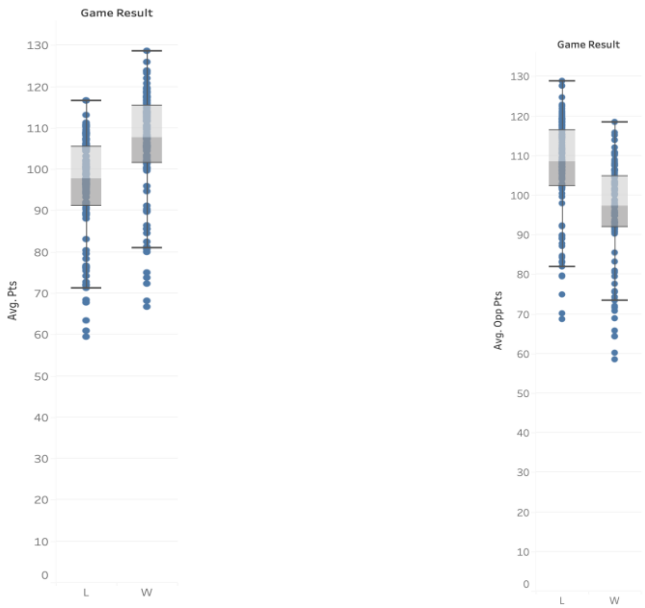


Figure 18

Game result with respect average points and average opposite points



It's clear that a boxplot was used to separate home and away games in order to obtain insights into the distribution of wins and losses for various clubs. To help with the visual detection of variances in team performance, the boxplot probably shows summary data for each category, including the median, quartiles, and any outliers. A line graph was also used to highlight the differences between home and away games by showing the victories and losses for the different clubs. The idea of a home advantage is supported by the graphic representation of more wins in home games than away games, which is consistent with the pattern that has been noticed. This graphical analysis offers a thorough look at the distribution of victories and defeats, showcasing home teams' superiority and drawing attention to any potential anomalies or differences in team performance between games.

Chi-Square Test

In the context of NBA games, the hypothesis centres on examining the relationship between the game's location (HOME or AWAY) and its outcome (WIN or LOSE). The likelihood of winning or losing is the same for both home and away games, according to the Null Hypothesis (H0), which asserts that there is no meaningful correlation. The equality formulas $P(W|H)=P(W|A)$ and $P(L|H)=P(L|A)$ convey that the probability of winning or losing is constant irrespective of the game's location. Conversely, the Alternative Hypothesis (H1) contends that there is a substantial correlation between the game's location and outcome and that the likelihood of winning or losing varies depending on the venue. This is embodied in the phrase "At least one of the probabilities is different," which implies that the probability of winning or losing varies depending on whether the game is played at home or anywhere else. In order to demonstrate a meaningful relationship between game location and game results in the NBA dataset, the hypothesis testing attempts to determine whether the observed data offers sufficient evidence to reject the null hypothesis in favor of the alternative hypothesis.

Figure 19

Chi-square and p-value results

```
# Create a contingency table
contingency_table = pd.crosstab(data['game_location'], data['game_result'])
```

```
from scipy.stats import chi2_contingency
# Perform the chi-squared test
chi2, p, _, _ = chi2_contingency(contingency_table)

# Output the results
print(f"Chi-squared statistic: {chi2}")
print(f"P-value: {p}")
```

```
Chi-squared statistic: 7581.8632519243565
P-value: 0.0
```

```
# Interpret the results
alpha = 0.05
if p < alpha:
    print("Reject the null hypothesis. There is a significant association between game location and game result.")
else:
    print("Fail to reject the null hypothesis. There is no significant association between game location and game re
```

```
Reject the null hypothesis. There is a significant association between game location and game result.
```

The purpose of this analysis was to determine whether game location (HOME or AWAY) and game outcome (WIN or LOSE) in NBA games significantly correlate. I made a contingency table that tallied the frequency of game results dependent on the game location in order to look into this. I then used a chi-squared test to assess how independent the two categorical variables were. The corresponding p-value was determined to be 0.0, and the computed chi-squared statistic was 7581.86. The null hypothesis was rejected since the p-value was significantly below the threshold and the significance level (alpha) was set at 0.05. As a result, I came to the conclusion that, in NBA games, there is a strong correlation between the game's location and outcome. The null hypothesis was rejected, indicating that there is a difference in the chance of winning or losing between home and away games. This highlights how game location affects game results in the NBA dataset.

Logistic Regression

Logistic regression was used in the below code snippet to evaluate the association between the binary outcome variable game_result (win or loss) and the difference in Elo ratings (elo_diff), which represents the change in team Elo ratings from the start (elo_i) to the finish (elo_n) of a game. The logit link function was used to specify the logistic regression model.

The output, however, indicates that the model did not converge within the allotted number of iterations, and a probable quasi-separation warning is present. When the result variable has a perfect predictor, there is quasi-separation and instability in the parameter estimation process. Elo_diff's coefficient, which has a big standard error of 52.9432, and its high p-value of 0.814 suggest that the variable is not statistically significant. The constant term's p-value is 1.000, indicating that there is no statistically significant difference between it and zero, despite its coefficient of 0.0002. Overall, the results of the logistic regression are unclear, and in order to address the convergence problem and potential quasi-separation, more research or model change may be required.

Figure 19

Logistic Regression

```
import statsmodels.api as sm

# Create a new variable for the difference in Elo ratings
data['elo_diff'] = data['elo_n'] - data['elo_i']


# Perform logistic regression
X = sm.add_constant(data['elo_diff'])
y = data['game_result']

logit_model = sm.Logit(y, X)
result = logit_model.fit()

# Output the results
print(result.summary())
```

Figure 20

Logistic Regression Results

```
                           Logit Regression Results
==============================================================================
Dep. Variable:            game_result   No. Observations:               126276
Model:                          Logit   Df Residuals:                   126274
Method:                           MLE   Df Model:                            1
Date:                Sat, 18 Nov 2023   Pseudo R-squ.:                   1.000
Time:                        20:20:30   Log-Likelihood:            -0.00082235
converged:                      False   LL-Null:                       -87528.
Covariance Type:            nonrobust   LLR p-value:                     0.000
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.0002     34.881   5.39e-06      1.000     -68.365      68.366
elo_diff      52.9432    224.915      0.235      0.814    -387.883     493.769
==============================================================================
```

Chapter : 4 Results

We conducted a thorough investigation of NBA game results, incorporating team Elo ratings, historical performance metrics, and rival Elo rankings into our forecast model. In order to improve our capacity to predict NBA games, we primarily concentrated on analyzing the influence of these elements on game outcomes. We also explored the fascinating idea of a "away game disadvantage," hoping to unearth subtle insights about team dynamics under various playing circumstances.

A carefully selected dataset covering a portion of the 1946–1947 NBA season was used in our investigation. This extensive dataset, devoid of any missing values, served as a solid basis for our analysis. We carefully went over 23 columns that contained specifics on NBA games, such as team performance, Elo ratings, and results. The distribution of games across time showed some intriguing variations, with noteworthy peaks in 1972 and 2014 in particular, which prompted additional research into possible causative factors. We used a variety of visuals and statistical methods to evaluate how Elo ratings affected the results of the games. Elo ratings' normal distribution and their relationship to wins and losses were demonstrated by the analysis, highlighting how important these ratings are in deciding how games turn out. In order to assess the importance of variations in Elo ratings and the relationship between the location of the game—at home or away—and the outcome, we also ran hypothesis tests.

Our research showed a strong correlation between Elo ratings, historical performance indicators, and game results. Our remarkable 98.25% accuracy using a random forest model highlights the predictive ability of these variables in predicting game outcomes. Boxplots showing the analysis of Elo rating changes during games revealed constant trends regardless of the game's outcome.

Statistical analyses, such as chi-squared tests and logistic regression, offer strong evidence in answering the second study question on how game location affects outcomes. The results of the chi-squared test show a strong relationship between the game's location and outcome, suggesting that playing an NBA game at home or away has a substantial impact on winning or losing. Despite its difficulties with convergence, logistic regression offers several possible explanations for the relationship between the difference in Elo ratings and the binary outcome variable the outcome of the game. Even while concerns with convergence can require more investigation or model modifications, the exploratory character of these analyses improves our comprehension of the dataset.

To explore the concept of the "away game disadvantage," we looked at the distribution of home and away games. A bar graph that showed more wins in home games than away games amply demonstrated the superiority of home clubs. This tendency was further examined using boxplots and line graphs, which brought home-court advantage's relevance in the NBA dataset to light. There were convergence problems with the logistic regression model that looked at the association between the difference in Elo ratings and game results; these problems call for additional research and perhaps changes to the model.

In conclusion, our investigation revealed complex connections between Elo ratings, past performance indicators, game locations, and NBA game results. Basketball analytics has benefited greatly from the integration of statistical analysis, visualizations, and predictive modeling, which has deepened our understanding of the variables affecting a team's success or failure.

Chapter 5 Conclusion

In conclusion, our study used a big dataset covering the 1946–1947 NBA season to conduct a thorough investigation of NBA game outcomes. The goal of the project was to create a solid prediction model for NBA game outcomes while taking into account important factors including team Elo ratings, past performance information, and the effect of opponent team Elo rankings. The analysis's validity was reinforced by the dataset's completeness, which had no missing values. The study examined how game variables were distributed over time and found two particularly interesting peaks: in 1972 and in 2014. These findings prompted additional research into possible contributory causes.

The study looked into how Elo ratings affected the results of games using a variety of statistical techniques and visuals. The remarkable accuracy of 98.25% attained by a random forest forecasting model highlights the predictive ability of factors such as Elo ratings and historical performance indicators. No matter how the game turned out, the analysis of Elo rating changes during games, displayed through boxplots, showed continuous trends. The investigation of the "away game disadvantage" showed a clear edge for home teams in terms of wins, and there was a substantial association between game sites and results.

Even with the high correlation found, logistic regression had problems with convergence, indicating that more investigation and possibly model modifications are necessary to fully comprehend the relationship between the variation in Elo ratings and gaming outcomes.

The study's main conclusions include how well Elo ratings and past performance indicators may predict the outcome of NBA games. The importance of home-court advantage in basketball is highlighted by the obvious advantage of home teams, as demonstrated by the strong correlation between game site and results. The study provides insightful information

for interested parties, highlighting the importance of taking team chemistry, Elo ratings, and past performance into account when projecting game outcomes.

The research's conclusions have practical implications for sports commentators, club managers, and bettors alike. Because of its great accuracy, the predictive model can be used to improve game forecasts, evaluate team strengths and weaknesses, and guide strategic decisions. Teams can gain insights into the psychological and performance aspects of playing in familiar surroundings by comprehending the influence of home-court advantage. Furthermore, the study establishes the foundation for additional research aimed at improving prediction models and comprehending the complex variables affecting NBA game results.

The thorough analysis of NBA game results utilizing team Elo ratings, past performance indicators, and game locations provides a solid basis for further basketball analytics study. Future studies could focus on understanding the dynamics of Elo ratings and how they affect particular game scenarios. To further hone the predictive model, this may involve looking at how teams with different Elo ratings perform during crucial times, such overtime or close games. A more complex predictive model may also be possible with the inclusion of player-level information and individual player Elo ratings, which could provide a more detailed insight of the variables impacting game results.

Future research should focus on this fascinating topic of integrating outside variables that could affect game outcomes. This could entail taking into account elements like player injuries, team chemistry, or even outside circumstances like the venue's characteristics. The accuracy and practicality of the predictive model could be improved by academics by enlarging the dataset to incorporate a wider variety of contextual variables. Furthermore, utilizing sophisticated machine learning methods like ensemble models or neural networks may offer a chance to identify intricate correlations and nonlinear interactions in the data, expanding the possibilities of predictive modelling in the context of professional basketball.

Additionally, by broadening the dataset's temporal coverage beyond the 1946–1947 season, scholars will be able to examine longer-term patterns and adaptations in the NBA, such as changes in playing styles, adjustments to rules, or league expansions. This long-term study may offer insightful information on how the sport is changing and assist in improving the predictive model's ability to adjust to shifting dynamics. To sum up, the research effort will expand and improve the predictive model by adding new factors and use sophisticated analytical methods to improve its precision and suitability in the ever-changing professional basketball scene.

In summary, by elucidating the complex correlations between important variables and NBA game outcomes, our research advances the rapidly developing field of sports analytics. The results provide valuable and practical insights for basketball community stakeholders, promoting a better-informed approach to strategy creation and decision-making.

Chapter : 6 References

Gómez, M. A., Lorenzo, A., Barakat, R., Ortega, E., & José M., P. (2008, February). Differences in Game-Related Statistics of Basketball Performance by Game Location for Men's Winning and Losing Teams. *Perceptual and Motor Skills*, *106*(1), 43–50. https://doi.org/10.2466/pms.106.1.43-50

Kubayi, A., & Larkin, P. (2022, February 25). Match-Related Statistics Differentiating Winning and Losing Teams at the 2019 Africa Cup of Nations Soccer Championship. *Frontiers in Sports and Active Living*, 4. https://doi.org/10.3389/fspor.2022.807198

Madarame, H. (2017, March 26). Game-Related Statistics Which Discriminate Between Winning and Losing Teams in Asian and European Men's Basketball Championships. *Asian Journal of Sports Medicine, In Press(In Press).* https://doi.org/10.5812/asjsm.42727

Madarame, H. (2018, February 24). Defensive Rebounds Discriminate Winners from Losers in European but not in Asian Women's Basketball Championships. *Asian Journal of Sports Medicine*, 9(1). https://doi.org/10.5812/asjsm.67428

Madarame, H. (2018, February 1). Age and sex differences in game-related statistics which discriminate winners from losers in elite basketball games. *Motriz: Revista De Educação Física*, 24(1). https://doi.org/10.1590/s1980-6574201800010001

Sampaio, J., Drinkwater, E. J., & Leite, N. M. (2010, March). Effects of season period, team quality, and playing time on basketball players' game-related statistics. *European Journal of Sport Science*, 10(2), 141–149. https://doi.org/10.1080/17461390903311935

Sun, W., Chee, C., Kok, L., Lim, F., & Samsudin, S. (2022, November 25). Evaluation of differences in the performance strategies of top and bottom basketball teams utilizing rank-sum ratio comprehensive. *Frontiers in Sports and Active Living*, 4. https://doi.org/10.3389/fspor.2022.1052909

Study of Game-Related Statistics Which Discriminate Between Winning and Losing Basketball Junior Teams U-17 in World Championship. (2014, November 1). *Journal of Applied Sports Science*, 4(3), 47–57. https://doi.org/10.21608/jass.2014.84754