Table of Contents

1. Data Overview

1.1 NBA Player Salaries (2022-23 Season)

The dataset for the Data 102 Final Project, titled "NBA Player Salaries for the 2022-2023 Season," is sourced from Kaggle. This dataset merges player per-game and advanced statistics for the NBA's 2022-23 season with player salary data, creating a comprehensive resource for understanding the performance and financial aspects of professional basketball players. The dataset is the result of web scraping player salary information from Hoopshype, and downloading traditional per-game and advanced statistics from Basketball Reference.

1.2 NBA data

The NBA Games dataset, with over 26,000 entries, details individual matches, including dates, teams, and a range of statistics like points scored and shooting percentages. Its 21 columns provide granular data for game-level analysis. Meanwhile, the NBA Team Ranking dataset logs over 210,000 entries on team standings over time, tracking wins, losses, and win percentages. Combined, these datasets offer a comprehensive toolkit for analyzing NBA game outcomes and team performance trends across multiple seasons, serving as a valuable resource for in-depth basketball analysis.

1.3 NBA All Stars 2000-2016

The NBA All-Star dataset provides a detailed account of NBA All-Star selections from 2000 to 2016. It encompasses various attributes of the players selected for the All-Star games. The data encompass the year of selection, player name, position (Pos), height (HT), weight (WT), team, selection type, NBA draft status, and nationality. The period of the dataset only overlaps 3 years with the main game dataset so we manually added the all-star rosters from 2017 to 2019.

3 Research Question 2

Will a previous-year NBA All-Star player cause the team's performance to improve in terms of win rate?

2.1 Introduction

2.1.1 Purpose

Importing a new player through trades, free agency, or the draft is a crucial aspect of building and maintaining a successful NBA team. Among these, the trade of an All-Star player will generally attract the most public attention. These All-Star players normally play pivotal roles in shifting the competitiveness of teams, impacting the strategies and dynamics of teams. Acquiring an All-Star player could have a profound influence on the franchise; for example, in 2018, LeBron James left the Cavaliers and signed with the Lakers. Ever since then, LeBron James has been the franchise player for the Lakers. Fans are excited about these trades, actively following the rumors during the offseason.

From the team's perspective, acquiring an All-Star player can serve as an immediate force on the court for the team. A famous example was the 2017-2019 Warriors. After a dominant regular season record in 2016, the team lost to the Cavaliers in the final. To everyone's surprise, the

Warriors acquired Kevin Durant, an All-Star player who had almost beaten them in the Western Conference final. Later, the Warriors won two consecutive NBA championships in 2017 and 2018 and barely lost to the Toronto Raptors in 2019 due to player injuries.

However, All-Star players are not the only factor contributing to the success of a team. Sometimes, acquiring an All-Star player could potentially break a team's chemistry, worsening the team's performance. Specifically, we are interested in the change in the team's performance in terms of win rate and whether acquiring a previous-year All-Star player has a positive impact on the team in general.

2.1.2 Model of choice

For this question, we decided to implement causal inference because it helps identify causality between acquiring a new All-Star player and the team's win rate. Since we calculated the difference in the win rate for each team every season and knew if a team imported a previous-year All-Star player at the beginning of the season, this is an observational study. Techniques like Outcome Regression and Inverse Propensity Weighting, which we learned in class, would be helpful. We exclude the method of matching due to the small sample size and difficulties in matching exact variables.

2.1.3 Data processing

Data processing is challenging because existing datasets do not contain all the information we need to conduct a causal inference experiment. For existing dataset, we use "games.csv", "games_details.csv", "teams.csv", "ranking.csv" and "teams.csv" in the Kaggle NBA datasets, with the external dataset "allstar.csv" we found for NBA All-Star rosters from 2000 to 2016. We added the roster information from 2017 to 2019 manually to the dataset.

We need to calculate the win rate for each team from 2013 to 2020. To achieve this, we create a function *find_winrate(df, team_id, season_id)*. The logic of this function is to find the maximum column "G" of a team in ranking.csv, which is the number of games played in a season, and then use "W_PCT", which is the win rate. Since the dataset records the win rate after every game, we used np.argmax and found the win rate in the last game in the season. We used for-loops to iterate through every season and every team. We calculated the difference between two win rates.

The most important function is *is_imported(team_id, season)*. This is the function that identifies whether the team acquires a new previous-year All-Star player. The logic of this function is to first find the last game of the season in games.csv, then use the game_id to locate all the players who played in that game in games_details.csv. We also repeated the procedure to the previous season stats. For each player in the roster, if that player is in the previous-year All-Star list and

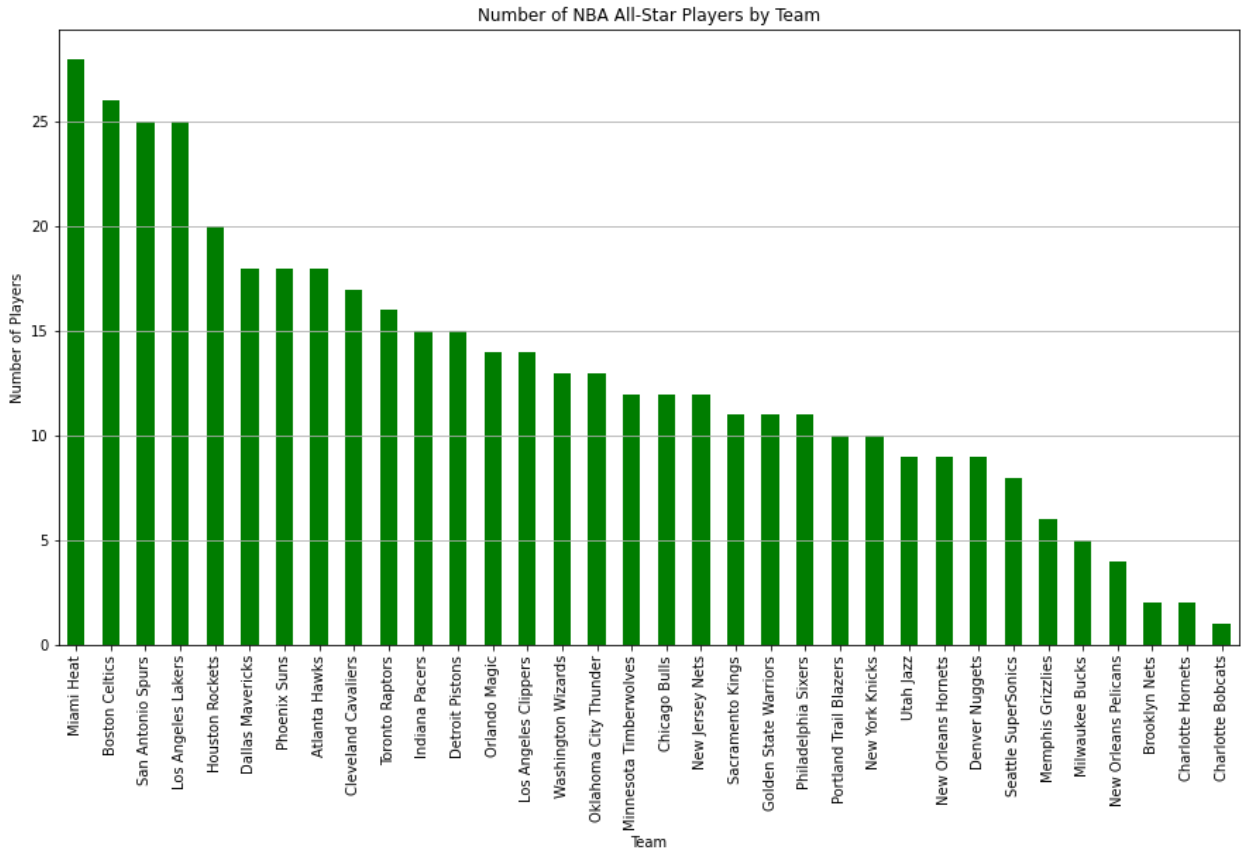
not in the previous season team roster, we marked the team in that season as 1 in the “treat” column, ‘0’ otherwise.

We used one-hot-encoding for the column “CONFERENCE” in teams.csv to make two columns “West” and “East”. We also counted the total number of All-Stars players in a team for every season as we thought it might be a confounder.

Following the above procedures, we generated the dataset contained columns “TEAM_ID”, “SEASON”, “treat”, “CURR_WINRATE”, “PREV_WINRATE”, “TEAM_NAME”, “DIFFERENCE”, “total_allstar”, “CONFERENCE”, “East”, “West”.

2.2 EDA

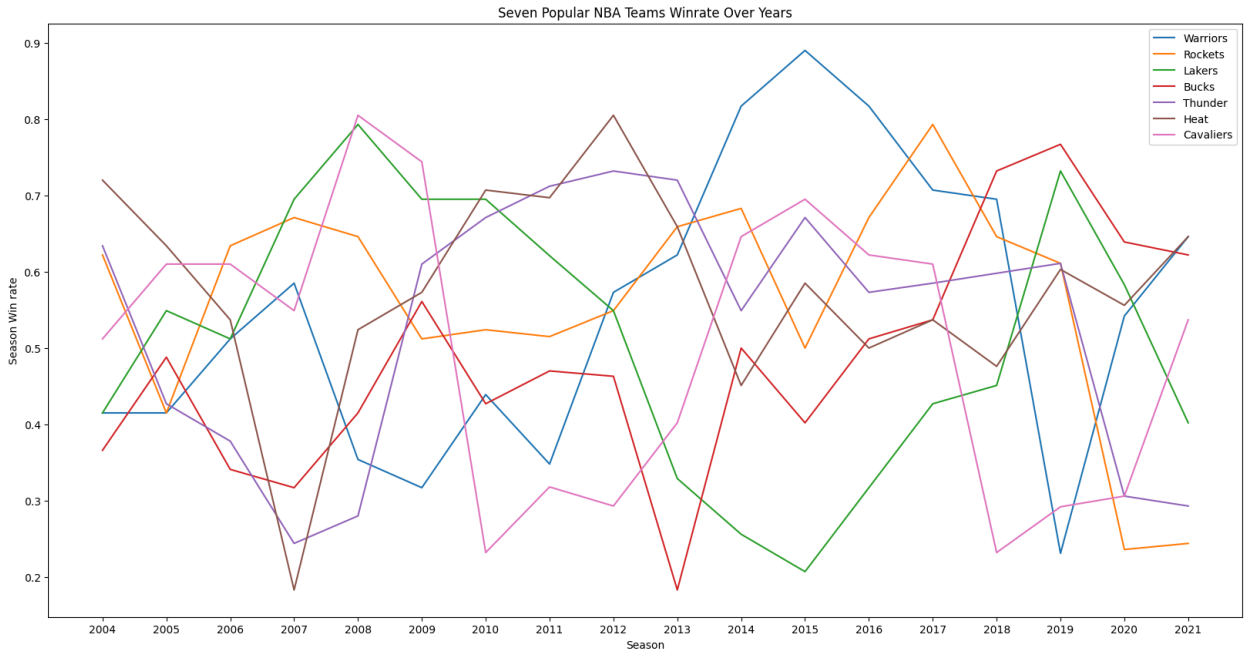
2.2.1 Bar Chart for number of NBA All-Star Players per Team



We created a bar chart illustrating the number of NBA All-Star players by team from 2000 to 2016. As seen in the graph, the total number of All-Star players for each team is unevenly distributed during this period. Additionally, a player can be selected multiple times as an

All-Star, reflecting the popularity and, to some extent, the power ranking of a team. A player must be beloved by the public or consistently perform well in their games to be selected as an All-Star. We believe such performance will impact the season win rate. This is also related to our hypothesis that the number of All-Star players in the current roster may influence the decision regarding whether a new All-Star player joins the team.

2.2.2 Line Plot for the Win Rate by Team

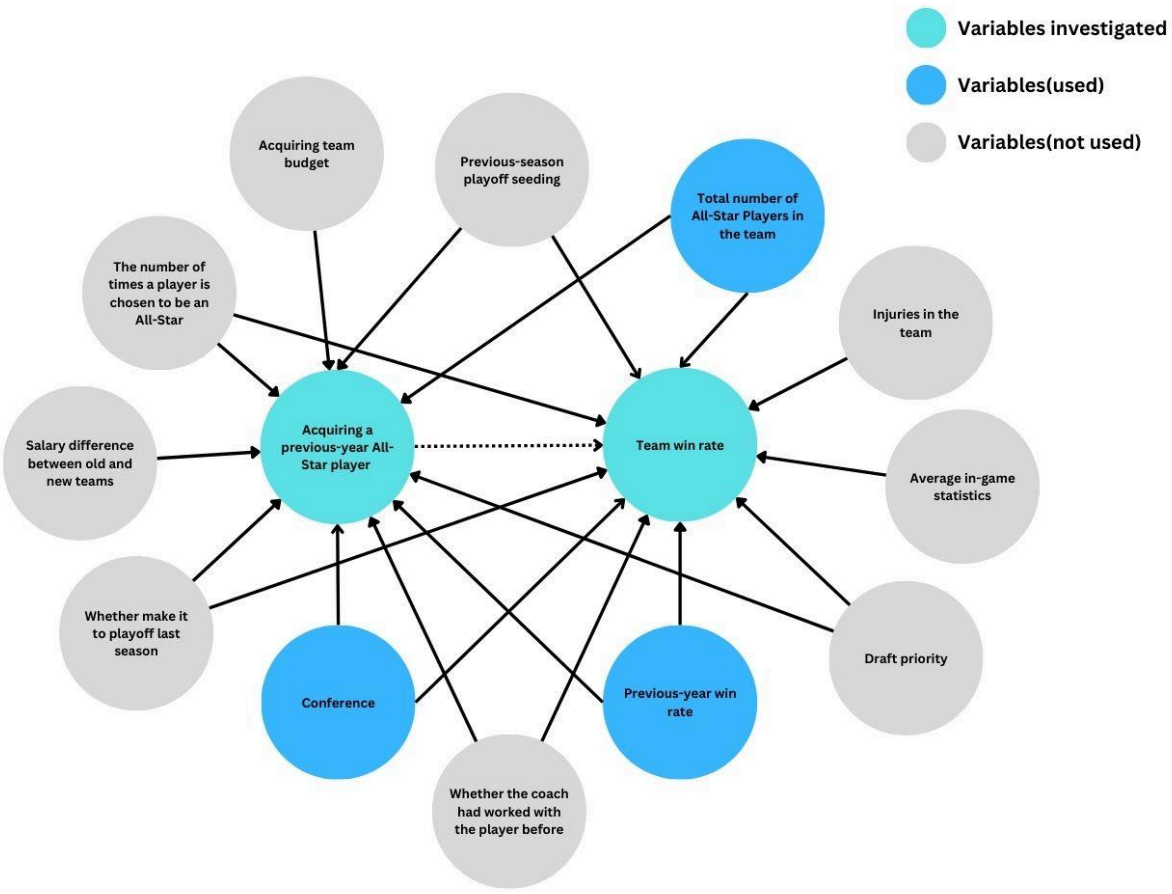


We created a line plot for seven popular NBA teams to illustrate their overall win rates in the regular season over time. Several interesting patterns emerge. After a dominant regular season with a 72-9 record (~89%), the Warriors acquired Kevin Durant in 2017-2018, a decision that helped them secure two championship titles. However, their regular season performance declined (still high, but ~70%). When the Heat lost their star player, LeBron James, in 2014, their regular season performance dropped significantly (from 66% to 45%), while the Cavaliers, who acquired LeBron James, saw a slight increase (from 65% to 70%). Noticeably, the Heat's win rate dropped significantly in the last season when LeBron played for them in 2013 (from 80% to 66%), which might contribute to the reason why they traded him or he left on his own will.

Based on our general knowledge about NBA market transfers, it appears that the graph shows either an increase or decrease in regular season win rate for a team acquiring an All-Star player. This makes our problem interesting and motivates us to continue the research.

2.3 Method

2.3.1 Assumption



(DAG)

In our method, the treatment is whether a team acquires a previous-year All-Star player. The outcome is the difference of team's performance, measured in terms of win rate. The unit is a

NBA team per season. We identified variables that have potential impact on team win rate are **average in-game statistics** and **injuries**. **Salary difference** between old and new teams and **acquiring team budget** will impact the decision of acquiring a previous-year All-Star player. For the confounding variables, we identified **the number of total all-star players** since an all-star player wants to join a champion-contending team to win the championship; the **conference** of the team because the competitiveness of each conference could be different sometimes. **Previous-season win rate** also impacts on the decision of acquiring an All-Star player and difference between two seasons' win rate. The above three confounders are what we used in the model.

We also identify confounders like **whether the team made it to the playoffs last season**, **last playoff seedings** since these teams attract all-stars who have never won before. **The number of times a player was chosen as an All-Star** (the value and strength of that player would be high), **draft priority** (teams with low win rate have higher priority and they could trade draft priority for all-star players), **coach worked with the All-Star players previously**(attract players and good chemistry), etc... Unfortunately, we could not find the dataset related to these topics and even we did, the work of merging these datasets is too complicated and time consuming for this project's purpose.

We also identified there exist colliders such as media attention. Whenever a team performs extremely well in the regular season or acquires a well-known All-Star Player would gain a lot of media attention. Unfortunately, it is hard to quantify the variables and we do not have enough data.

Thus, we assume the unfoundedness by identifying three confounding variables as all confounding variables, and there are no unobserved confounding variables.

We first calculated Simple Difference in the observed group to get a general idea.

2.3.2 Outcome Regression

Assuming there are no unobserved confounding variables, we can use outcome regression to adjust the influence of confounders. By unconfoundedness, we can fit a linear regression with treatment variables and all confounding variables we used in the problem, using difference of win rate as outcome. In addition, assume this linear model correctly describes the interaction between the variables. The estimated coefficient of treatment from OLS, tau, will be an unbiased estimate of the ATE.

2.4 Results

The simple difference in the observed group means.

```
#Compute the Simple Difference in Observed group means (SDO) for this observational data.

sdo = sum(df[df['treat'] == 1]['DIFFERENCE']) / len(df[df['treat'] == 1]) - sum(df[df['treat'] == 0]['DIFFERENCE'])/
sdo
```

```
0.01586229946524064
```

Fitting all of our variables into a linear regression model.

```
### all confounders
linear_model = fit_OLS_model(df, 'DIFFERENCE', ['treat','PREV_WINRATE','total_allstar','East','West'])
print(linear_model.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:             DIFFERENCE   R-squared:                       0.274
Model:                            OLS   Adj. R-squared:                  0.260
Method:                 Least Squares   F-statistic:                     19.33
Date:                Sun, 10 Dec 2023   Prob (F-statistic):           1.64e-13
Time:                        13:06:11   Log-Likelihood:                 158.85
No. Observations:                 210   AIC:                            -307.7
Df Residuals:                     205   BIC:                            -291.0
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
treat          0.0594      0.024      2.486      0.014       0.012       0.106
PREV_WINRATE  -0.5267      0.061     -8.687      0.000      -0.646      -0.407
total_allstar  0.0260      0.010      2.495      0.013       0.005       0.047
East           0.2308      0.029      8.044      0.000       0.174       0.287
West           0.2399      0.031      7.802      0.000       0.179       0.301
==============================================================================
Omnibus:                        7.369   Durbin-Watson:                   1.980
Prob(Omnibus):                  0.025   Jarque-Bera (JB):                7.109
Skew:                          -0.417   Prob(JB):                       0.0286
Kurtosis:                       3.342   Cond. No.                         12.1
==============================================================================

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

The model's coefficient for "treat"(~6%) is larger to the result in the simple difference(~1.5%)

For every variable, we are 95% confident that the coefficient interval does not include 0, making the result significant.

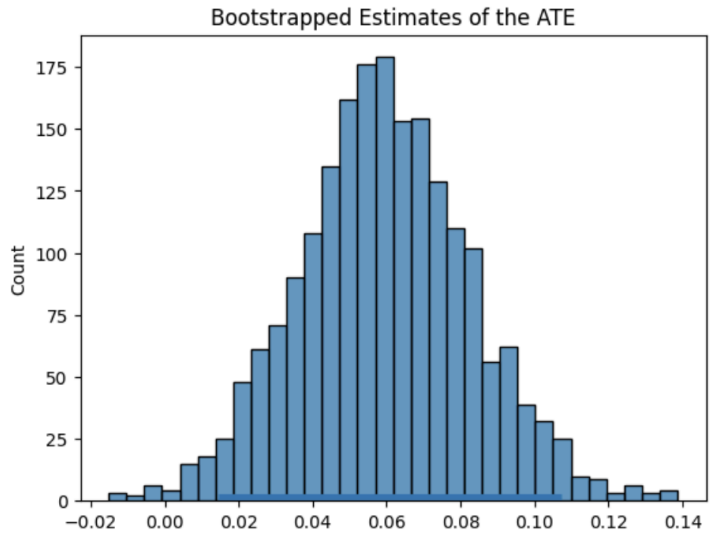
To test the impact of randomness in the model, we also tried to use bootstrap through a frequentist's perspective.

```
ates = get_bootstrapped_ate(df, 2000)
confidence_interval = [np.percentile(ates, 2.5),
                       np.percentile(ates, 97.5)]
print(f"Our 95% confidence interval ranges from {(confidence_interval[0])} to {(confidence_interval[1])}")
```

```
Our 95% confidence interval ranges from 0.014444216763565722 to 0.10732870083308062
```

```
sns.histplot(ates)
plt.hlines(1, confidence_interval[0], confidence_interval[1], linewidth=5)
plt.title("Bootstrapped Estimates of the ATE")
```

```
Text(0.5, 1.0, 'Bootstrapped Estimates of the ATE')
```



We can see that the 95% confidence interval captures the value in simple difference and does not include 0, making the result significant.

Even though the simple difference indicates an insignificant improvement in terms of win rate(~1.5%), our model suggests an approximate 6% improvement in win rate. Both Bayesian and Frequentist's perspectives agree with the result. However, the magnitude of this improvement is not too big, indicating there exists many other factors that impact the win rate of a team.

2.5 Discussion

The limitation of our method is that we have to assume there are no unobserved confounding variables, but they obviously exist in the real world. As we indicated in our assumptions and

directed acyclic graph (DAG), there are numerous confounding variables and colliders that we did not account for in our model due to its complexity.

When we calculated the simple difference value, we were disappointed to see such a small improvement in the win rate, considering the tremendous effort we put into processing the data. The later analysis using outcome regression and bootstrapping provided a better result and demonstrated why simple difference estimation is a flawed approach.

In the future, we can definitely utilize more datasets involving coaches, player injuries, playoff seedings, team budgets, etc. These are the confounding variables that we identified in the above section that would impact both the decision of a team acquiring an All-Star and the team's win rate. Additionally, we should reconsider the usage of 2020 data because of the Covid pandemic.

We are confident in our conclusion because of the significant result, but we understand that we only considered three unconfounding variables in our model. The unconfoundedness assumption does not hold in our cases. Adding more confounding variables may change our result.

2.6 Conclusion

In the research, we assume that every confounding variable is observed. We found that acquiring a previous-year All-Star player will increase the team's performance by 6% in terms of the win rate. Our finding is based on the years 2013 to 2020, which is not highly generalizable, considering the NBA's first All-Star game took place in 1951. This limited time span resulted from constraints in our dataset. However, it does reveal patterns and impacts of recent trades involving All-Stars.

In the end, the win rate improvement is lower than we expected. If we ignore the imperfect aspects of our approach and datasets, based on the results, we want to suggest that NBA teams focus on developing young, new rookies instead of acquiring well-known All-Star players. Unless you aim to improve your team immediately without budget considerations, developing new players or acquiring decent but not All-Star level players could be more beneficial for the team.

Citation

NBA Player Salaries (2022-2023 Season)
https://www.kaggle.com/datasets/jamiewelsh2/nba-player-salaries-2022-23-season/data?select=nba_2022-23_all_stats_with_salary.csv

NBA games data
https://www.kaggle.com/datasets/nathanlauga/nba-games/data?select=games.csv

NBA All Stars 2000-2016
https://data.world/gmoney/nba-all-stars-2000-2016