Data 102 Final Project
Written Report

Fall 2023

Bill Tian, Michael Li, Weiqian Peng, Yanyu Chan

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.

2. Research Question 1

Can you predict players' salaries using players' statistics?

2.1 Introduction

2.1.1 Purpose

In the competitive landscape of the NBA, understanding the factors that influence a player's salary is crucial for team management, sports analysts, and fans alike. This project aims to clarify the financial valuation of NBA players by developing a predictive model that links their salaries with various on-court performance metrics and other relevant factors. Utilizing a rich dataset from the NBA's 2022-23 season, we delve into the complex interplay between athletic performance and financial rewards.

Our methodology involves comprehensive data preprocessing to refine the dataset, followed by careful feature engineering to capture the most relevant aspects influencing a player's salary. Through statistical modeling, we aim to not only predict salaries with a degree of accuracy but also to uncover the key attributes that command higher financial compensation in the NBA. This project serves as a bridge between data analytics and sports management, providing insights that can guide contract negotiations, team building, and player development strategies. By the end of this report, we will have explored the nuances of our model's construction, its predictive capabilities, and the implications of our findings in the broader context of professional basketball economics.

2.1.2 Model of choice

In our project, we have chosen to utilize both Generalized Linear Models (GLM) and nonparametric models to predict NBA players' salaries, leveraging the strengths of each to address the complexity of our data.

GLM are favored for their interpretability and ability to model different response variable distributions, crucial for dealing with the continuous and varied nature of salary data. GLMs excel in providing insights into linear relationships between variables, making them ideal for initial analysis and understanding direct effects of features on salaries.

On the other hand, nonparametric models offer unparalleled flexibility by not assuming a specific functional form between predictors and response. This makes them adept at capturing complex, nonlinear relationships that GLMs might overlook, essential in a dataset where such intricate patterns can exist due to the diverse statistics influencing player salaries.

By integrating both GLM and nonparametric models, we aim to build a comprehensive and robust analytical framework. GLM helps us establish baseline relationships and interpret the direct impact of predictors, while nonparametric models delve deeper into uncovering hidden patterns and nonlinear dynamics. This dual approach ensures a more thorough and nuanced understanding of the factors that drive NBA players' salaries.

2.1.3 Data processing

The dataset that we used was acquired from Kaggle, a well-known platform for data projects, and downloaded the csv file as ‘nba_2022-23_all_stats_with_salary.csv’. This dataset is a census that includes all NBA players for the 2022-23 season. As previously mentioned, the data in this dataset utilized the player salary information from Hoopshype and advanced statistics from Basketball Reference. There is no indication that any group was systematically excluded as it included every single player. These statistics are publicly available and lots of statistics are usually used to determine the value and training of a player, so these NBA players should know that this information is being collected. One row represents an individual NBA player in the 2022-23 season. Since this dataset is a census, there should be no selection bias. As this information is from official websites and the NBA statistics, measurement error and convenience sampling should have been kept to a minimum. The dataset was not modified for differential privacy as previously stated that the dataset includes public sports data. Certain features that we wished we had were whether or not the player was injured and for how many games the players were injured. In addition, we wished we had the player’s biographical data as this may be a factor in determining a player’s salaries. There does not appear to be any missing data.

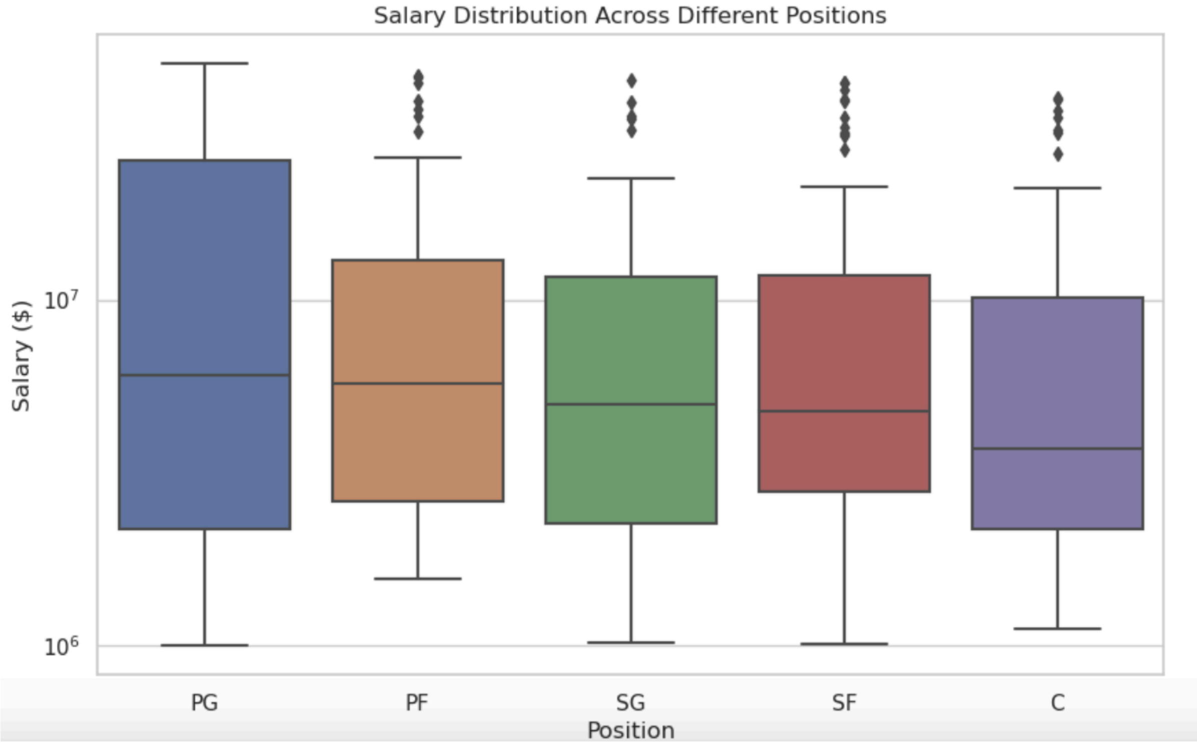
We filtered out basketball players who make less than $1000000. We chose this amount because the minimum starting salary for the year 2022-2023 was this amount. The salaries below this value were usually rookies who just debuted that year and played only a few games. For instance, the minimum salary was $5849 and the individual only played one game which makes this player an outlier as the stats that he got in this game does not reflect his overall salary.
In addition, there were seven players that played two positions. Since the number of players who play two positions account for less than 2% of the total number of players, we decided to choose the primary position that these players play. For instance, the players who play both PG and SG positions, we converted their positions to only include PG and removed SG from their positions. This allowed our analysis to only include the main 5 positions of NBA players: PG, PF, SG, SF, and C. We believe this is a reasonable move to make as we believe that positions play a role in determining salaries and having only the 5 positions that players play provides the emphasis needed on positions. We also added a new feature of games started divided by games played. We believe this is a better feature than looking at the two features individually because certain players may not play a lot during the season midway as they get injured, but they are still star

players for their team. We used one-hot-encoding for the column "Position" to make five columns "PG", "PF", "SG", "SF" and "C".

The features we selected include 'PG', 'PF', 'SG', 'SF', 'C', 'Age', 'FGA', 'TOV%', 'PTS', 'PER', 'VORP', 'FG%', 'USG%', 'GS/GP' to predict 'Salary'.

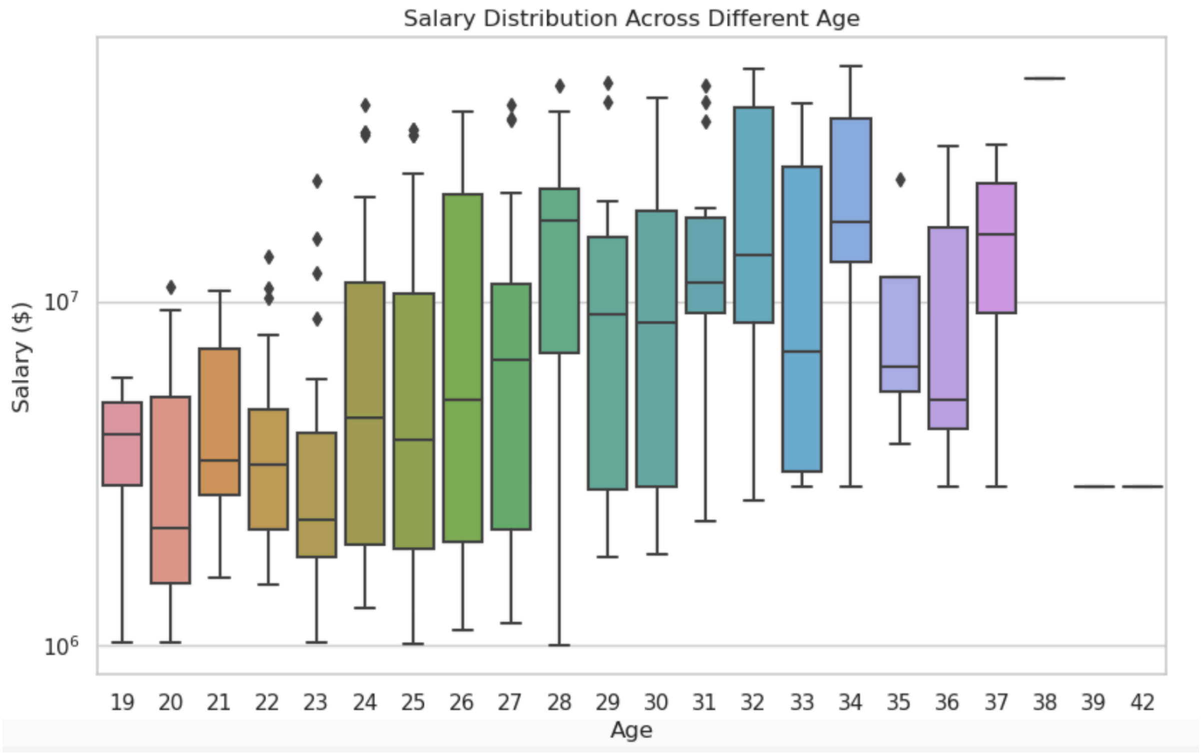
2.2 EDA

2.2.1 Visualizations of relevant features



From the graph "Salary Distribution Across Different Positions", we can infer that there are slight variations in salary distributions among the different positions: Center (C), Power Forward (PF), Point Guard (PG), Small Forward (SF), and Shooting Guard (SG).

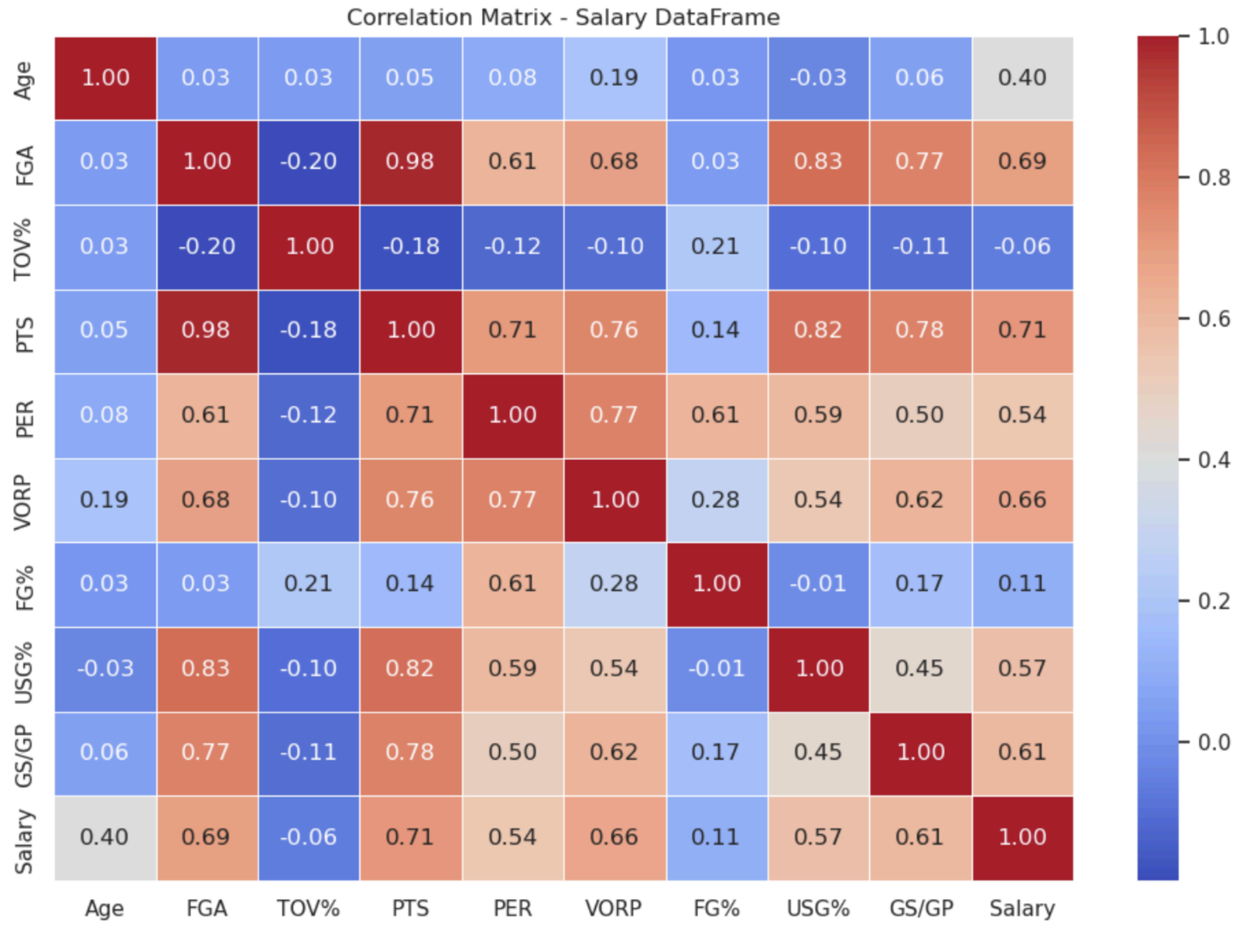
PG and PF have higher median salaries compared to the other positions, which is evident from the position of the median line in their respective boxes. There are several outliers for each position, except for PG, visible as individual points above the upper whiskers of the boxplot. These outliers represent players with exceptionally high salaries compared to their peers.



From the plot “Salary Distribution Across Different Ages”. We can see that there is a general trend of increasing salaries from ages 19 to a peak around 25-30, after which the median salary tends to slightly decrease. The variation in salaries, as indicated by the spread of the boxes and the length of the whiskers, is substantial for each age group, with a tendency to increase during the early to mid-career ages.

Given the above two plots, we can conclude that playing positions and age of the players have a potential impact on players’ salaries.

2.2.2 Heatmap of Other Features


Correlation Matrix - Salary DataFrame

	Age	FGA	TOV%	PTS	PER	VORP	FG%	USG%	GS/GP	Salary
Age	1.00	0.03	0.03	0.05	0.08	0.19	0.03	-0.03	0.06	0.40
FGA	0.03	1.00	-0.20	0.98	0.61	0.68	0.03	0.83	0.77	0.69
TOV%	0.03	-0.20	1.00	-0.18	-0.12	-0.10	0.21	-0.10	-0.11	-0.06
PTS	0.05	0.98	-0.18	1.00	0.71	0.76	0.14	0.82	0.78	0.71
PER	0.08	0.61	-0.12	0.71	1.00	0.77	0.61	0.59	0.50	0.54
VORP	0.19	0.68	-0.10	0.76	0.77	1.00	0.28	0.54	0.62	0.66
FG%	0.03	0.03	0.21	0.14	0.61	0.28	1.00	-0.01	0.17	0.11
USG%	-0.03	0.83	-0.10	0.82	0.59	0.54	-0.01	1.00	0.45	0.57
GS/GP	0.06	0.77	-0.11	0.78	0.50	0.62	0.17	0.45	1.00	0.61
Salary	0.40	0.69	-0.06	0.71	0.54	0.66	0.11	0.57	0.61	1.00

We compared the correlation between each feature and we selected these following ten features to construct a heatmap to obtain more intuitive data furtherly.

The heatmap analysis of NBA player salaries and performance metrics yields several key insights. Primarily, 'PTS' or points per game stands out with a strong positive correlation of 0.71 with salary, indicating that players who score more tend to earn more. This underscores the premium placed on scoring in the NBA. 'FGA', or field goal attempts, is also closely related with a correlation of 0.69, reinforcing the value of offensive involvement in determining a player's salary. 'VORP', or Value Over Replacement Player, with a correlation of 0.66, is another significant metric. It reflects a player's overall contribution to the team, suggesting that players who perform well above the average are highly valued and thus better compensated. In contrast, metrics like 'FG%', which measures shooting efficiency, shows a relatively weak correlation of 0.11 with salary. This suggests that while efficiency is beneficial, it is not as heavily weighted in salary considerations as the ability to score and the overall impact on the team's performance. The heat map tells us that in the NBA, the ability to score and a player's integral role in their team's strategy are paramount in influencing salary.

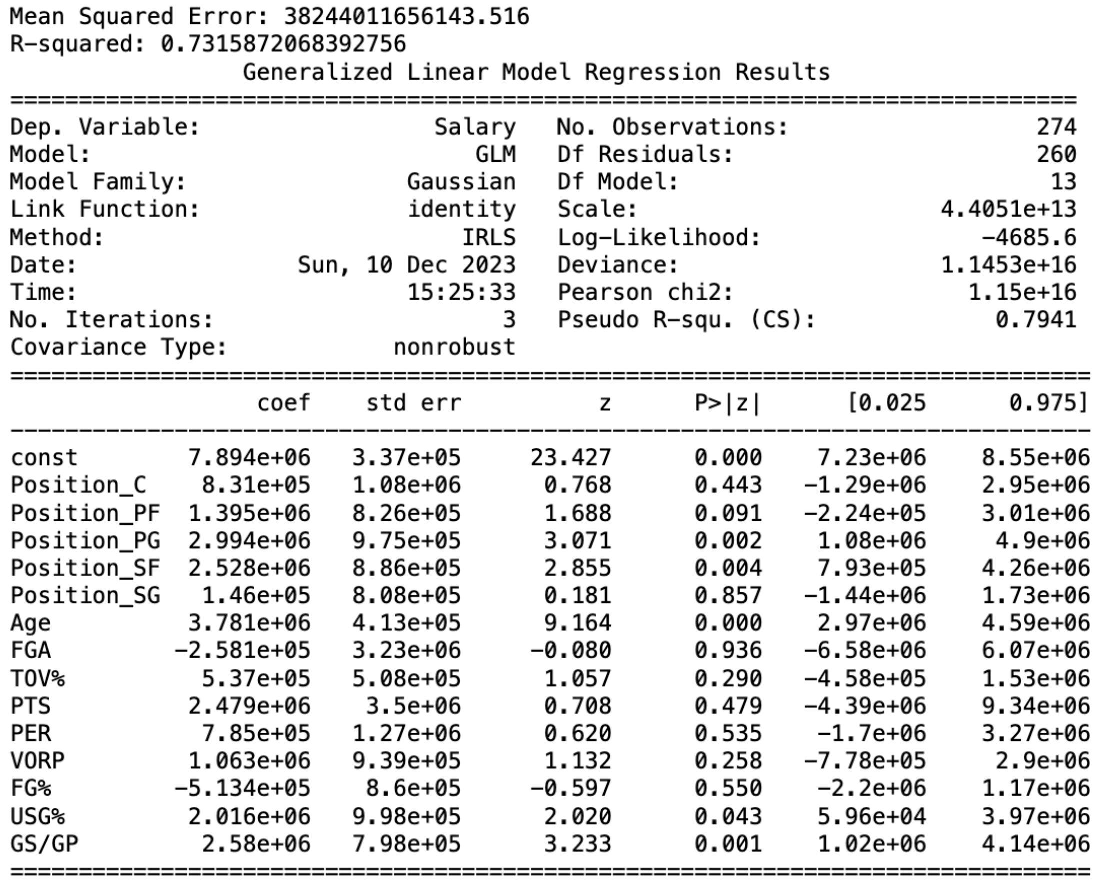
2.3 Method

In our model, the observed variables include measurable player statistics like points scored, field goals attempted, and age. We're trying to estimate "hidden" variables that aren't directly observable, like a player's potential future performance or their market value beyond what their current stats suggest.

For GLM, we used a Gaussian distribution, assuming that player salaries are normally distributed around an average similar to a bell curve. Gaussian distribution is a good fit when the data clusters around a central mean and decreases symmetrically on both sides, which is often the case with salary data. This distribution helps in modeling the data accurately, allowing us to make more reliable inferences about the relationship between player attributes and their salaries. And the non-parametric method was a random forest with not much assumptions. The choice of a Random Forest model for our data is rooted in its versatility with the kind of data we have. Random forest does not need the data to follow a specific statistical distribution, which is great for the complex and varied nature of player statistics. It can handle the intricacies and interactions between different player attributes effectively.

2.4 Results

2.4.1 Frequentist GLM results



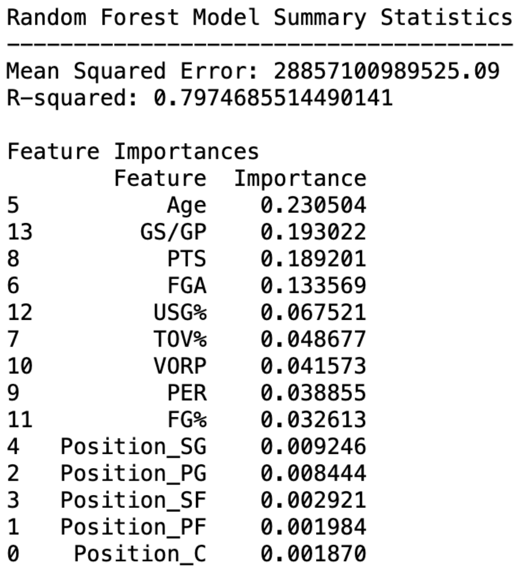
```
Mean Squared Error: 38244011656143.516
R-squared: 0.7315872068392756
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:                 Salary   No. Observations:                  274
Model:                            GLM   Df Residuals:                      260
Model Family:                Gaussian   Df Model:                           13
Link Function:               identity   Scale:                      4.4051e+13
Method:                          IRLS   Log-Likelihood:                -4685.6
Date:                Sun, 10 Dec 2023   Deviance:                   1.1453e+16
Time:                        15:25:33   Pearson chi2:                 1.15e+16
No. Iterations:                     3   Pseudo R-squ. (CS):             0.7941
Covariance Type:            nonrobust
===============================================================================
                  coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------
const        7.894e+06   3.37e+05     23.427      0.000    7.23e+06    8.55e+06
Position_C    8.31e+05   1.08e+06      0.768      0.443   -1.29e+06    2.95e+06
Position_PF  1.395e+06   8.26e+05      1.688      0.091   -2.24e+05    3.01e+06
Position_PG  2.994e+06   9.75e+05      3.071      0.002    1.08e+06     4.9e+06
Position_SF  2.528e+06   8.86e+05      2.855      0.004    7.93e+05    4.26e+06
Position_SG   1.46e+05   8.08e+05      0.181      0.857   -1.44e+06    1.73e+06
Age          3.781e+06   4.13e+05      9.164      0.000    2.97e+06    4.59e+06
FGA         -2.581e+05   3.23e+06     -0.080      0.936   -6.58e+06    6.07e+06
TOV%          5.37e+05   5.08e+05      1.057      0.290   -4.58e+05    1.53e+06
PTS          2.479e+06    3.5e+06      0.708      0.479   -4.39e+06    9.34e+06
PER           7.85e+05   1.27e+06      0.620      0.535    -1.7e+06    3.27e+06
VORP         1.063e+06   9.39e+05      1.132      0.258   -7.78e+05     2.9e+06
FG%         -5.134e+05    8.6e+05     -0.597      0.550    -2.2e+06    1.17e+06
USG%         2.016e+06   9.98e+05      2.020      0.043    5.96e+04    3.97e+06
GS/GP         2.58e+06   7.98e+05      3.233      0.001    1.02e+06    4.14e+06
===============================================================================
```

The Generalized Linear Model (GLM) we've developed to predict NBA player salaries is fairly robust, explaining over 73% of the variance in salaries, as indicated by an R-squared value of 0.7316. Key positions like Point Guards and Small Forwards, as well as metrics such as 'USG%' (Usage Percentage) and 'GS/GP' (Games Started per Games Played), are found to have a significant positive impact on salary levels, highlighting the value placed on players who have a central role in gameplay and team strategies.

On the flip side, more traditional statistical measures, specifically 'PTS' (Points) and 'PER' (Player Efficiency Rating), don't show the expected significant correlation with salary. Intriguingly, 'FGA' (Field Goal Attempts) has a significant but negative association with salary, suggesting that merely taking more shots isn't necessarily rewarded with higher pay.

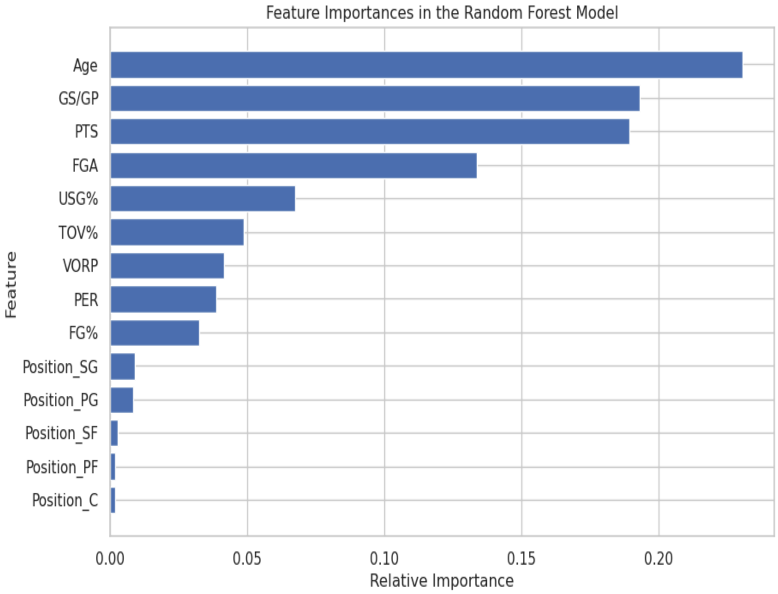
The model does reveal areas of uncertainty, particularly where the 95% confidence intervals for certain variables, including several positions and performance stats, encompass zero. This points to a lack of definitive evidence on how these factors influence salaries. Given these uncertainties and the chance of omitted variable bias, it's clear that further refinement of the model is necessary. Additional data and alternative modeling techniques may yield more precise insights into salary determinants, ultimately enhancing decision-making processes for NBA team management and player contract negotiations.

2.4.2 Random Forest Model results



```
Random Forest Model Summary Statistics
--------------------------------------
Mean Squared Error: 28857100989525.09
R-squared: 0.7974685514490141

Feature Importances
        Feature  Importance
5           Age    0.230504
13        GS/GP    0.193022
8           PTS    0.189201
6           FGA    0.133569
12         USG%    0.067521
7          TOV%    0.048677
10         VORP    0.041573
9           PER    0.038855
11          FG%    0.032613
4   Position_SG    0.009246
2   Position_PG    0.008444
3   Position_SF    0.002921
1   Position_PF    0.001984
0    Position_C    0.001870
```



Mean squared error is defined as the absolute squared difference between the observed actual outcomes and the predicted outcomes of our model. Our random forest model has a fairly big mean squared error, but since we are trying to predict salary which is in the millions of dollars, it is difficult to interpret if this mean square error is a good or bad error. R-squared is defined as how close the data is to the fitted regression line. We got a 0.79747 as our R-squared which is a relatively high number. This indicates that 79.7% of the variance in the dependent variable, salary, can be explained by the model.

We also calculated the feature importances of each feature to display how much each feature contributes to the model's predictions. The feature with the highest importance is age with an importance score of 0.230504. This suggests that a player's age is a significant predictor in the model, which could be due to factors such as peak performance years or experience in the league. We did see some level of this in our EDA where salary seems to be positively correlated with age until a certain point and salary seemed to decrease slightly for the oldest players.

The next features with the highest importance are games started / games played and points. They have an importance score of 0.193022 and 0.189201, respectively. These are likely key performance indicators in basketball. If the proportion of games started / games played, it is likely that you are a major player for the team, leading to higher salaries. This is the same with points, as teams would like players who score more points, which can lead to a higher win rate.

Each of the five basketball positions seem to have a very low feature importance score. The position of SG has the highest feature importance of 0.009246. Followed by, PG with 0.008444, SF with 0.002921, PF with 0.001984, and C with 0.00187 importance score. This indicates that positions play a very small role in the model's predictions. As we saw in the box plot diagram from our EDA section, the salaries of the players split into the different positions are very similar and there was no significant difference between the positions.

2.5 Discussion

Our analysis utilized GLM and Random Forest models to predict NBA player salaries. The Random Forest model outperformed the GLM, as indicated by its higher R-squared value, suggesting it is more adept at capturing the complexities of NBA salary structures due to its ability to model non-linear relationships. However, applying these models to future datasets requires caution due to the evolving nature of the sport and its economics.

The GLM provided insights into linear relationships but showed limitations in capturing the full complexity of salary determinations. The Random Forest model, while more comprehensive, could be less interpretable and risks overfitting. Both models have their strengths, but the Random Forest's flexibility makes it more suitable for this dataset.

Incorporating additional data, such as player injuries, marketability, and team performance, could improve model accuracy. The uncertainty in our results is moderate, stemming from the intricate factors influencing player salaries and potential dataset limitations.

In summary, while both models offer valuable insights into NBA salary structures, their inherent limitations and the moderate level of uncertainty in the results suggest that model refinements and additional data could enhance future predictions and analyses in this dynamic field.

2.6 Conclusion

Our model's results indicate that certain player statistics significantly influence NBA salaries, like age, PTS. This suggests teams should prioritize these aspects in player development and contract negotiations. However, the analysis is limited by its focus on quantifiable metrics, potentially overlooking some factors like player marketability. On the other hand, the team's funds and background are also one of the important reasons affecting players' salaries. In the future research we could explore these qualitative aspects. We recommend NBA teams and agents use these insights for strategic decisions, rational and reasonable analysis of player value focusing on enhancing player skills that correlate strongly with salary growth.

3 Research Question 2

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

3.1 Introduction

3.1.1 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.

3.1.2 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.

3.1.3 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.

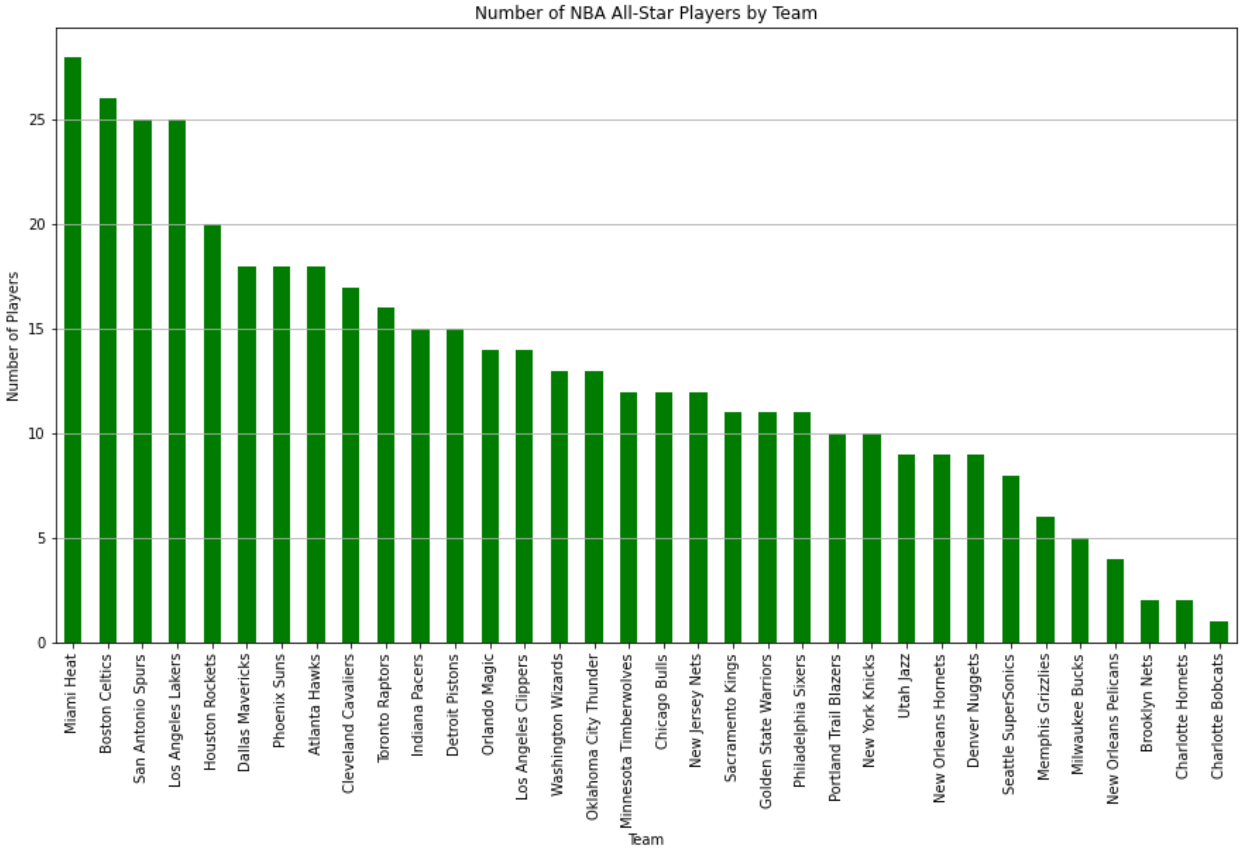
The most important function is *is_imported(team_id, season)*. This is the function that identifies whether the team acquires a new previous-year All-Star player. The logic of this function is to first find the last game of the season in games.csv, then use the game_id to locate all the players who played 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”.

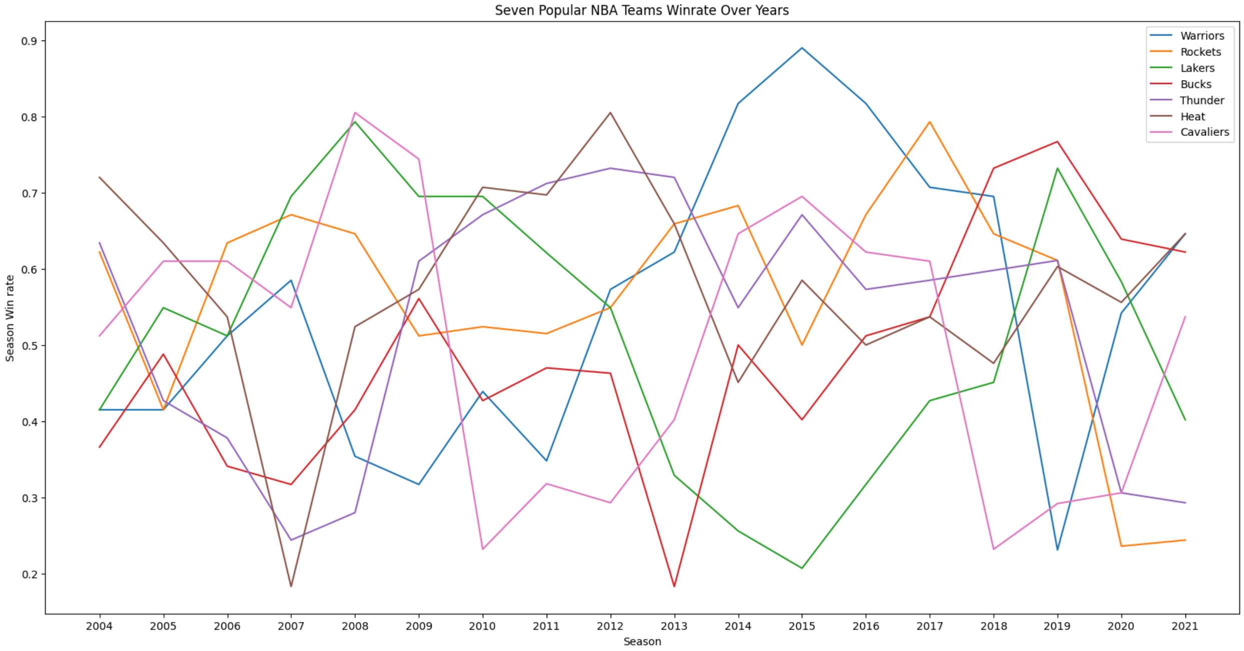
3.2 EDA

3.2.1 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.

3.2.2 Line Plot for the Win Rate by Team

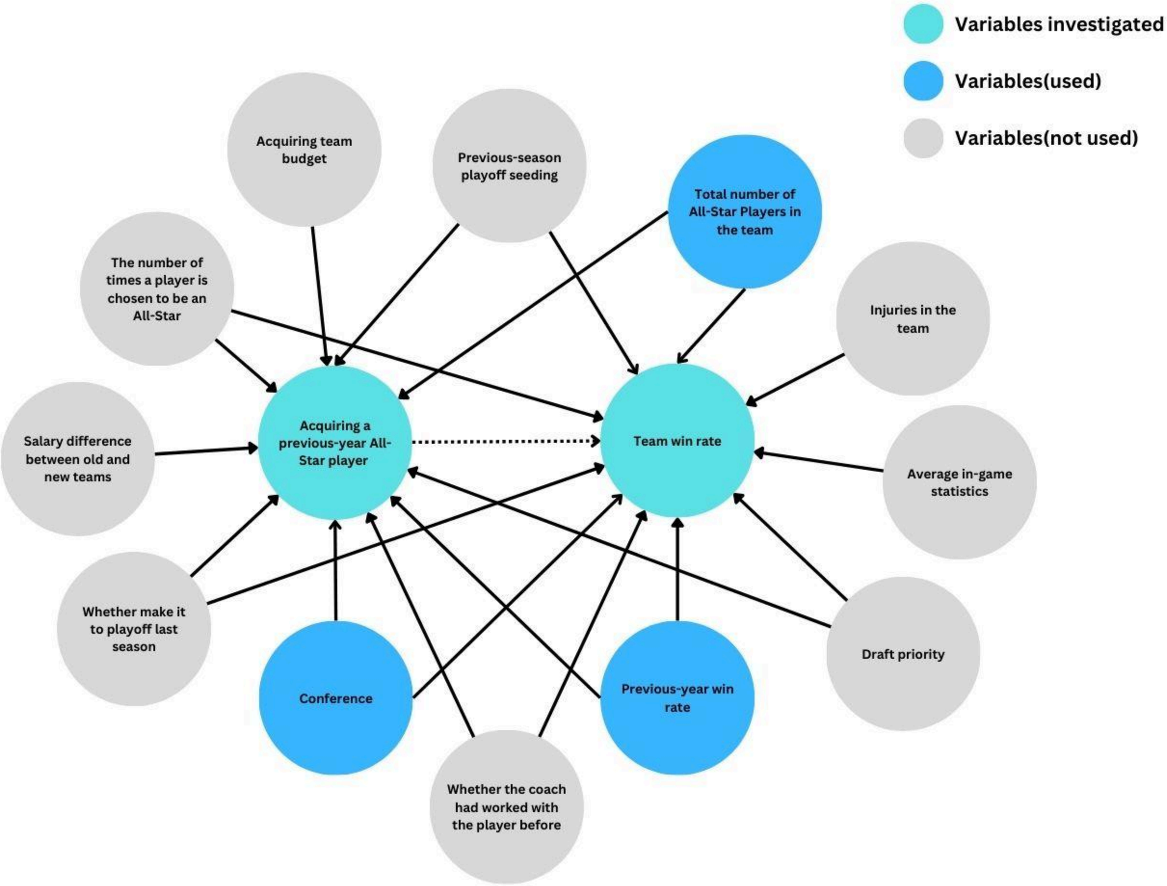


We created a line plot for seven popular NBA teams to illustrate their overall win rates in the regular season over time. Several interesting patterns emerge. After a dominant regular season with a 72-9 record (~89%), the Warriors acquired Kevin Durant in 2017-2018, a decision that helped them secure 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.

3.3 Method

3.3.1 Assumption


Variables investigated
Variables(used)
Variables(not used)
Acquiring team budget
Previous-season playoff seeding
Total number of All-Star Players in the team
The number of times a player is chosen to be an All-Star
Injuries in the team
Salary difference between old and new teams
Acquiring a previous-year All-Star player
Team win rate
Average in-game statistics
Whether make it to playoff last season
Draft priority
Conference
Previous-year win rate
Whether the coach had worked with the player before

(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.

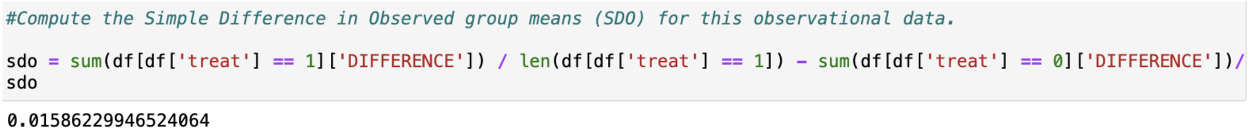
We first calculated Simple Difference in the observed group to get a general idea.

3.3.2 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.

3.4 Results

The simple difference in the observed group means.

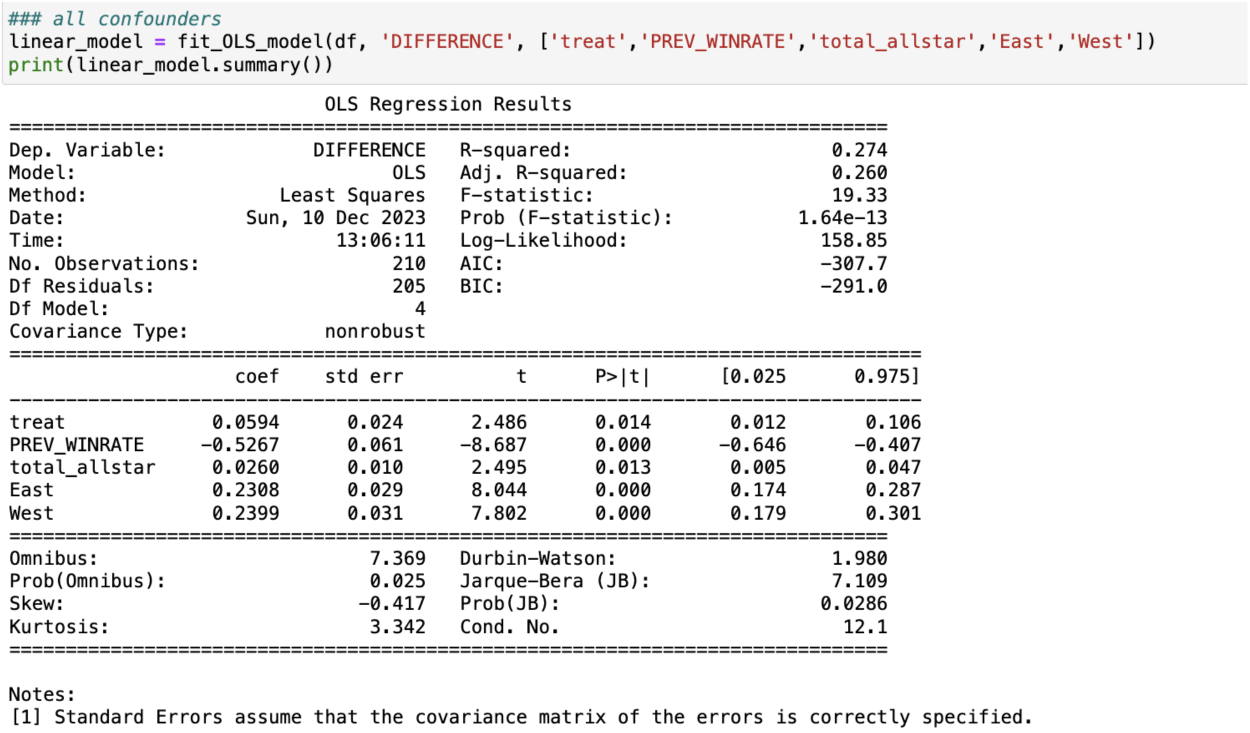


```
#Compute the Simple Difference in Observed group means (SDO) for this observational data.

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())
```

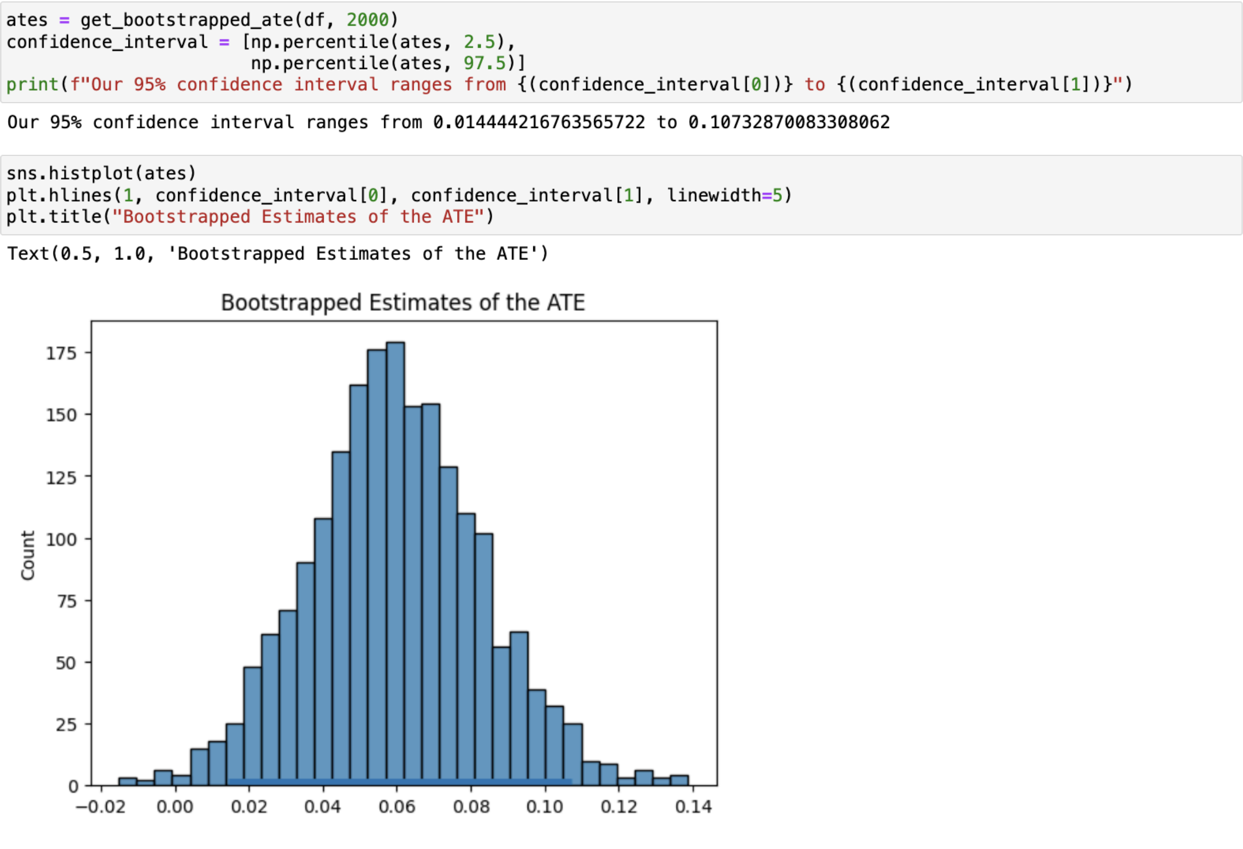
```
                            OLS Regression Results                            
==============================================================================
Dep. Variable:             DIFFERENCE   R-squared:                       0.274
Model:                            OLS   Adj. R-squared:                  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.

3.5 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.

3.6 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