Final Project 102

Parth Deepak Shisode, Julia Pastis, Jessica Wang, Purbasha Majee

Data Overview

How was your data generated? Is it a sample or census?

- Our data was collected through the NBA stats website, which hosts all the data from each game that has been played in the NBA. The games dataset contains statistics of every NBA game played since 2003. Games_details holds statistics for each player for each game played since 2003. Players hold information for which player played for what team for each season. Ranking holds rankings for each team for each season since 2003. Teams hold information for each team in the NBA. Our data is a census as for each data frame we have statistics about every team, player, and game from 2003 to 2022.

- Our data set was not modified for differential privacy.

- We have more than enough features to answer our research questions; there are actually additional statistics present within the dataset that could be used to explore additional questions on top of what we already intended to answer

- In the games_details dataset, it was found that about 16% of the data had missing values from important columns and the games data set about .03% had important columns with missing data. Since 16% is a relatively significant percentage of missing data, we decided that we would fill the columns about with its respective mode as based as we saw it would not alter the distribution of the columns in any significant way. We used this method for both data sets because of how we approached our analysis, we needed to approach how we dealt with missing values for both datasets in the same way.

- Dropping missing values was the only cleaning we needed to do within our data. We did not need to perform any pre-processing.

Research Questions

Option A: Multiple Hypothesis Testing / Decision Making

Question, Algorithm, and Modeling Choices

Our Approach: For our study, we focused on analyzing the impact of a star player's presence on various team performance metrics in NBA games. We employed A/B testing as our primary method, given its effectiveness in comparing two versions (with and without the star player) under similar conditions. Our hypotheses were centered around the assumption that the team's performance metrics would be statistically similar, regardless of the star player's participation.

Assumptions

Key Assumptions: Our analysis hinges on the assumption that each test's results are independent, due to the random nature of shuffling and sampling in A/B testing. While this might not always hold true (e.g., team dynamics can be complex and interdependent), it simplifies our approach and allows for clearer interpretation of results.

Implementation and Statement of Results

Implementation and Findings: We correctly implemented A/B testing, utilizing statistical software tools. Our findings indicate that, for certain teams, the presence of the star player significantly alters team performance metrics, as evidenced by p-values below the 0.05 significance level. This was observed across various metrics like FGP, BLKs, TOs, ASTs, 3PT%, and STLs.

Interpretation of Results

Understanding Our Findings: The initial results suggested a significant impact of the star player's presence. However, upon applying the Bonferroni and Benjamini-Hochberg correction methods, these significant findings disappeared, indicating that our initial results might have been influenced by the multiple hypothesis testing issue.

Conclusion and Future Directions

In conclusion, our study suggests that the presence of a star player does not significantly impact the team's performance across the tested metrics when considering the corrections for multiple

hypothesis testing. This finding emphasizes the need for caution when interpreting statistical results in sports analytics.

For future work, I propose extending this analysis to multiple seasons and possibly incorporating more nuanced metrics that capture a player's impact beyond traditional statistics. Additionally, a more powerful computational approach could enable us to increase the number of repetitions in our A/B testing, potentially yielding more reliable results.

We acknowledge the limitation of our computational capacity, which restricted the number of repetitions in our tests. Also, we consciously avoided p-hacking by employing rigorous threshold correction methods, ensuring our results weren't just artifacts of multiple testing or selective reporting.

Option C: Prediction with GLMs and nonparametric methods

Question, Algorithm, and Modeling Choices

We chose logistic regression for our GLM approach and a random forest model for nonparametric analysis. Our choice was influenced by the nature of our data and the binary outcome we are predicting – game wins or losses. Logistic regression is apt for binary data, while the random forest model's ability to handle various stats and binary outcomes makes it a suitable nonparametric choice.

In the GLM, we selected points, defensive rebounds, assists, and blocks per game based on their correlation and impact on the Akaike Information Criterion (AIC). For the random forest model, its nonparametric nature did not necessitate a manual feature selection process.

Assumptions

We divided our data based on whether the team was playing at home or away, as this distinction showed different correlation coefficients, aligning with the general understanding that home teams often have an advantage.

We assumed binary nature for wins and losses in logistic regression. However, we did not explore this assumption in depth for the Negative Binomial and Poisson models, which are not ideally suited for binary outcomes, as evidenced by their results.

Implementation and Statement of Results

We used accuracy and False Discovery Rate (FDP) to evaluate our models. The Random Forest model showed higher accuracy (~80-81%) and lower FDP compared to the Logistic model.

The Logistic Regression model, despite its lower accuracy compared to the Random Forest model, provided valuable insights through its coefficients and was more interpretable.

Interpretation of Results

The GLM models, especially logistic regression, were more interpretable but less accurate. The Random Forest model, despite being a black-box model, provided higher accuracy and seems more suitable for prediction tasks.

The uncertainty in the logistic regression coefficients, as evidenced by their standard errors, indicates a reasonable level of confidence in the model's predictions.

Conclusion

Real-World Applications: Based on our results, the Random Forest model seems more apt for predicting game outcomes in real-world scenarios, although its 'black-box' nature might limit interpretability. The logistic regression model, while less accurate, offers valuable insights into which factors significantly influence game outcomes.

Exploring additional contextual factors like player injuries, rest days, and individual player stats could improve model accuracy. Also, integrating a Bayesian approach could provide a different perspective, particularly in handling prior information and uncertainty.

Our study faces limitations due to the inherent unpredictability in sports and potential data noise. Future work should focus on addressing these limitations and exploring additional methods to improve the prediction accuracy of sports outcomes.

Option B: Bayesian Hierarchical Modeling Research Question

The research aims to estimate the win percentage of home and away teams in basketball over several seasons using Bayesian Hierarchical Modeling.

A Beta distribution is chosen as a prior for the win percentages, based on empirical data and domain knowledge suggesting a higher win rate for home teams in the NBA.

The choice of a Beta distribution is appropriate because it is commonly used for probabilities and rates. The parameters α and β are derived from the mean and variance found in existing literature.

The observed variables are the win/loss outcomes for each game (home and away), modeled using a Bernoulli process with the win percentage as the probability parameter.

Assumptions

The independence of games: Each game's outcome is considered independent of others, which might not always hold true due to factors like team form, injuries, etc.

Normal distribution for the prior: This assumes that the distribution of win percentages over seasons follows a normal distribution, which may not capture potential non-linear trends or anomalies in the data.

Stability of home-court advantage over time: The model assumes that the home-court advantage does not significantly change across seasons.

Implementation and Statement of Results

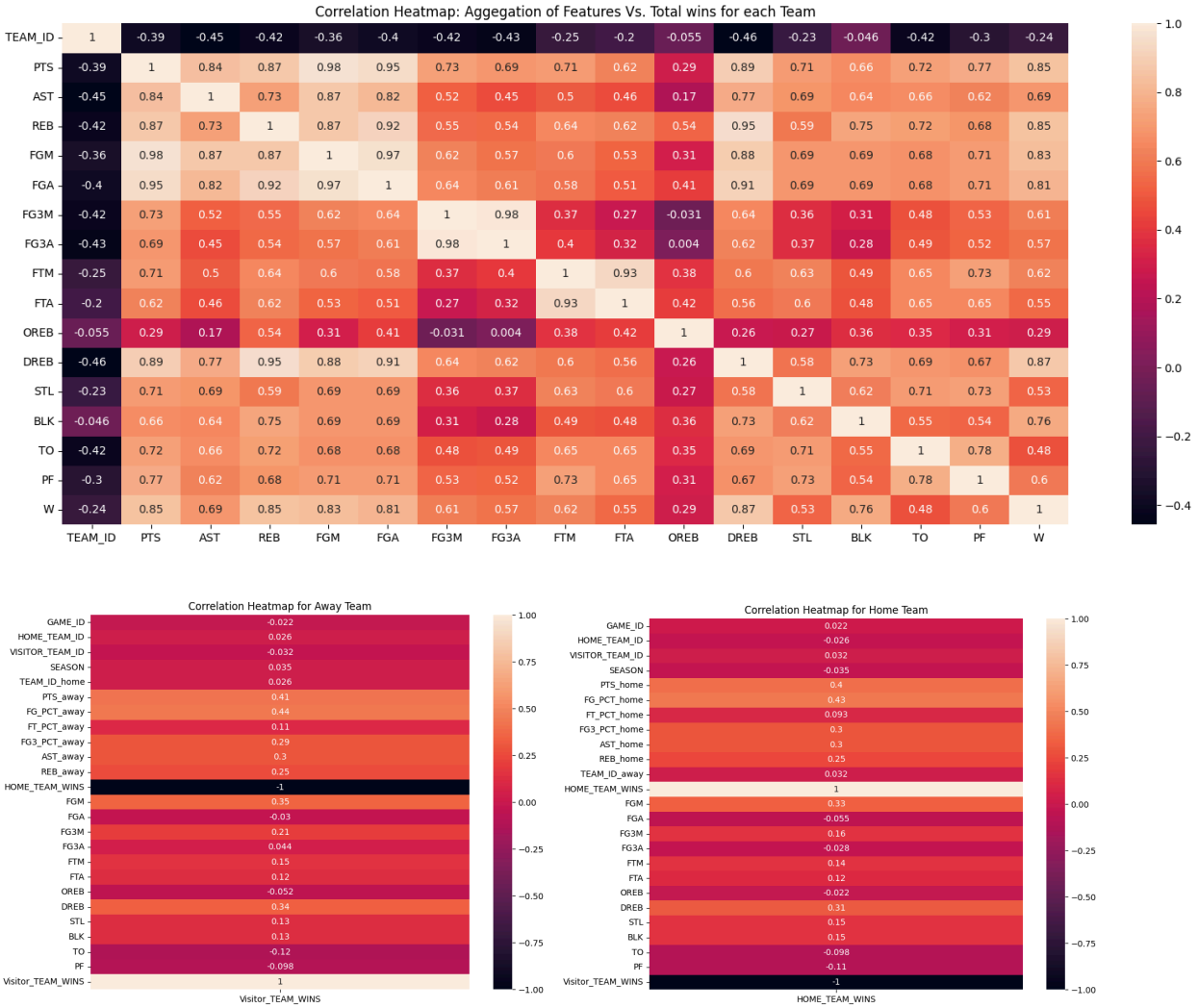
The model is implemented correctly, using Bayesian methods and PyMC for estimation.Home teams are estimated to have a 58% chance of winning, slightly lower than historical data suggests.Away teams have a 41% chance of winning, aligning with historical data.These results indicate a potential shift in the traditional home-court advantage.

Interpretation of Results

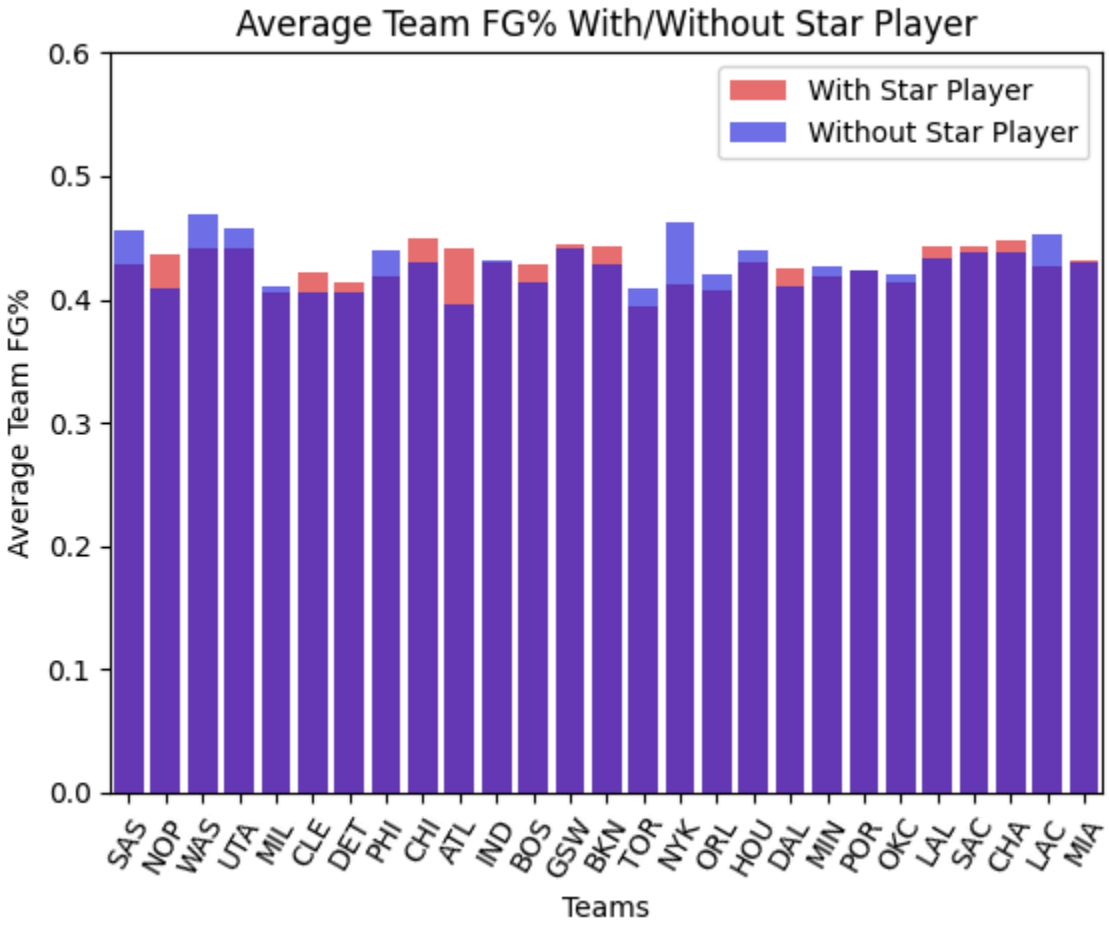
The results suggest a slight deviation from historical trends in home-court advantage, which could be due to various unmodeled factors like changes in team strategies, player fitness, or other external factors.

Conclusion

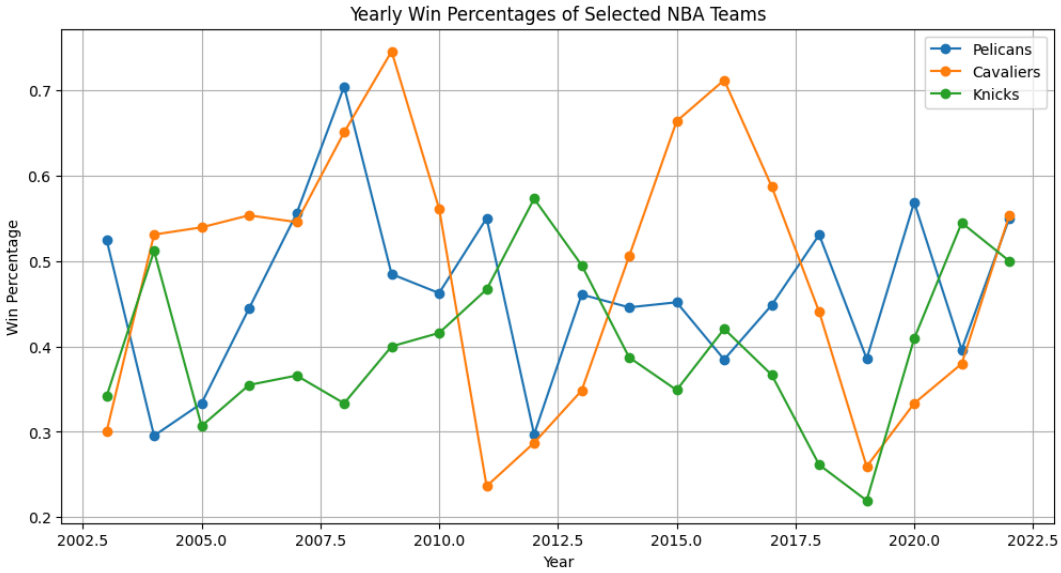
This analysis can inform team strategies and betting practices.
The model assumes game independence and consistent home-court advantage over time, which might not be accurate.
Investigating the temporal dynamics of home-court advantage and incorporating more nuanced data (like player stats, team form, etc.) could provide deeper insights. Further research using more dynamic models and richer datasets is recommended to fully understand the trends in basketball game outcomes.



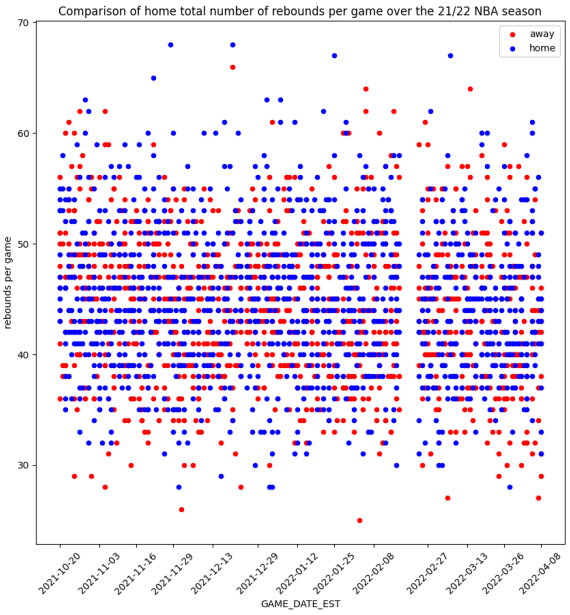
The above graph relates to our second research question. The graphs above are correlation heatmaps, which gives us information on how correlated each feature is to one another. This is relevant information, as it will help us choose which features would be good to use when building our regression model. From the visualization above, we see points scored, assists, and defensive rebounds, and blocks per game for each team is correlated with the number of wins for each team are good features to use as they have a high correlation coefficient. Our observations are consistent with the visualizations below, where the features specified above seem to have high correlation with whether or not the team would win the game.

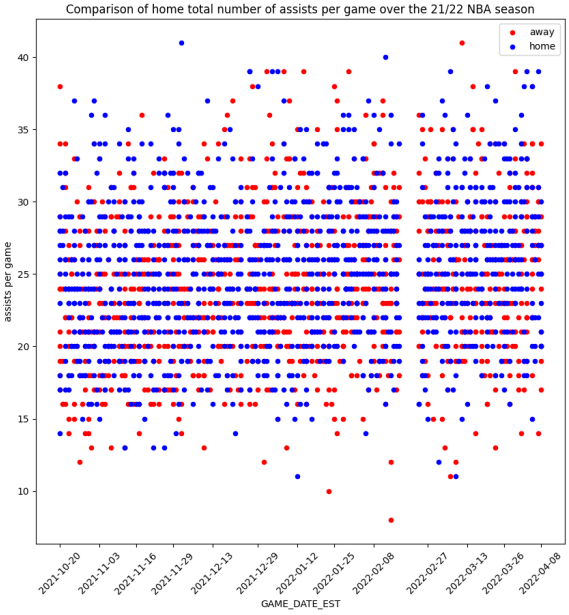


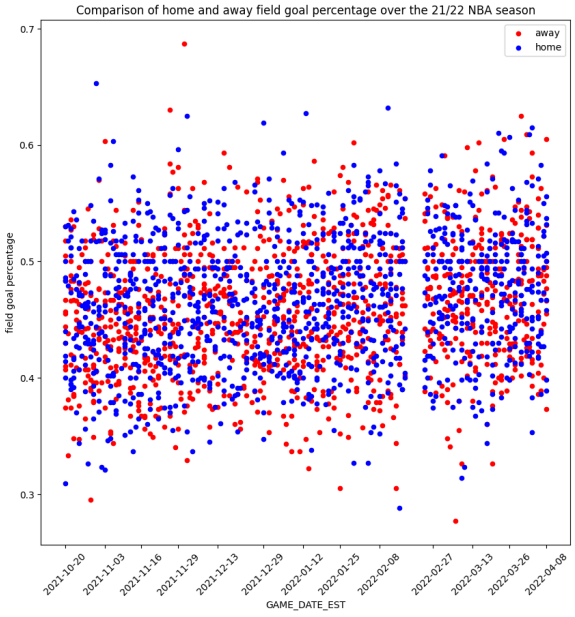
We can see here that at first glance, there doesn't appear to be a drastic difference in the average team field goal percentage when the star player plays or not. In terms of relevance to our research question, we now know that using field goal percent as a metric may not make it clear as to whether the team has a higher or lower shooting accuracy when the star players are in the game. Of course, star players are only identified by having the highest PPG on their team that season; other aspects of their game such as assisting teammates, offensive rebounding, etc. may also play a role in average field goal percentages changing when they are in the game. We will proceed with this knowledge.



We examine the win percentages of the New Orleans Pelicans, Cleveland Cavaliers, and New York Knicks. The line graph illustrates the win percentages from 2002 to 2022, indicating fluctuations that do not immediately correlate with the absence of star players. This suggests that additional factors may influence game outcomes and warrants a deeper statistical exploration.To align our EDA with the research question on the impact of star players, we will conduct a more granular analysis. This will involve examining game-level data to identify trends related to player absences. We also acknowledge that the graph does not account for external factors such as injuries, trades, or schedule density, which could affect team performance. The next phase of our EDA will compare other performance metrics, like player efficiency ratings, with win percentages during periods when star players were absent. This comprehensive approach aims to isolate the effect of star players on team success, providing a more robust understanding of basketball as a team sport where collective efforts often dictate game outcomes.







We were interested in how different stats change over the course of the season, if at all, and in comparing home vs. away teams for these. This is in reference to our second research question: Can we use random forests and logistic regression to predict which team will win out of a certain head to head match up. From the graphs above, we can see that there isn't much of a change over the course of a season in terms of these variables. And the trends don't change much between the home or away team. That's helpful in picking variables for our regression. Since there isn't much of a change over time in these, then we're probably accurate throughout the regular season with our estimation, no matter for which time period we use our model to predict. We're also going to be fairly accurate for the home or away team since they both seem pretty similar. So, we can stick with just one model with the most correlated variables, and can use that model to predict any point in the regular season.

Option A: Multiple Hypothesis Testing / Decision Making

Hypotheses:

- Our hypotheses all surround searching for differences in player performance when a team's "star player" plays in a game:

1. For a team, the distribution of the team's average FGP (field goal percentage) is the same in games where the star player is present as compared to games where the star player is not present; the difference in sample is due to random chance
2. for a team, the distribution of the team's average BLKs (number of blocks) is the same in games where the star player is present as compared to games where the star player is not present; the difference in sample is due to random chance
3. for a team, the distribution of the team's average TOs (number of turnovers) is the same in games where the star player is present as compared to games where the star player is not present; the difference in sample is due to random chance
4. for a team, the distribution of the team's average ASTs (number of assists) is the same in games where the star player is present as compared to games where the star player is not present; the difference in sample is due to random chance
5. for a team, the distribution of the team's average 3PT% (three-point field goal percentage) is the same in games where the star player is present as compared to games where the star player is not present; the difference in sample is due to random chance
6. for a team, the distribution of the team's average STLs (number of steals) is the same in games where the star player is present as compared to games where the star player is not present; the difference in sample is due to random chance

- We are making an assumption that the tests' results are independent from each other due to the random nature of the shuffling & sampling procedure in A/B testing

- We chose to test hypotheses using a wide range of performance-based statistics. We want to be able to test for improvements in several aspects of the game when the "star player" is present, not just scoring.
- Each of the above hypotheses has an alternative hypothesis, all six of which have a similar format. Here is the example of the alternative counterpart for hypothesis #1 from above:

  For a team, the distribution of the team's average FGP (field goal percentage) is higher in games where the star player is present as compared to games where the star player is not present.

- Power is not a relevant parameter to calculate for the way we're approaching this task; this is not a classification problem.

- Our group chose to go with A/B testing because it allows us to take advantage of the fact that whether the star player is in the game or not is a binary variable. This method's purpose is to allow us to determine whether two sets of data are sampled from the same distribution; in our case, A/B testing allows us to determine whether performance data is similarly distributed based on the presence of the star player. Additionally, A/B testing also allowed us easy interpretation of our results.

Threshold Correction Methods

- Our group decided to utilize the Bonferroni and Benjamini-Hochberg (B-H) correction procedures. Bonferroni correction controls Family-Wise Error Rate (FWER) and B-H correction controls False Discovery Rate (FDR)
- Both of these were (Bonferroni for controlling FWER and B-H for controlling FDP) made fewer discoveries than naive. Bonferroni was far more conservative though, rejecting everything since we were doing this test across so many teams. Since this is multiple hypothesis testing we want to avoid a naive threshold right away. Using just a naive threshold would add the typical type 1 error up for however many tests we're doing. We wanted to pick a threshold with less type 1 error than that. Second, we have all the data, so LORD isn't needed in this case. Finally, between B-H and Bonferroni, we concluded B-H was the best choice for us. Making a type 1 error (or thinking that a star player had a statistically significant effect on a stat when they really didn't) when measured with the tradeoff of a type 2 error (or thinking that a star player doesn't have a statistically significant effect when they really do) doesn't seem like a horrible tradeoff. Therefore, we were willing to sacrifice some type 1 error for some more type 2 accuracy. Therefore, we picked the slightly less conservative BH method to establish our cutoff.

Results

```
Teams Which Meet Naive Threshold
Field Goal %: ['NOP']
3-Pt Field Goal %: ['CLE']
Blocks per Game: ['SAS', 'GSW']
Assists per Game: ['BKN']
Steals per Game: ['SAS', 'GSW']
Turnovers per Game: ['BKN', 'ORL', 'LAC']
```

- The above teams, for each of these performance statistics, hold p-values below the significance level of 0.05. In other words, for these teams the presence of the star player leads to distributions of performance statistics which are higher.

- We chose BH and Bonferroni. For BH, we're controlling FDP by adapting the threshold cutoff value to the specific p-values in that test. For Bonferroni, we're controlling for FWER by ensuring the chance of a false positive isn't higher than the original alpha value.

Discussion

```
Benjamini-Hochberg (B-H) Threshold Correction: p-values
Field Goal %: 0
3-Pt Field Goal %: 0
Blocks per Game: 0
Assists per Game: 0.0025
Steals per Game: 0
Turnovers per Game: 0.016
```

```
Teams Which Meet B-H Corrected Threshold
Field Goal %: []
3-Pt Field Goal %: []
Blocks per Game: []
Assists per Game: []
Steals per Game: []
Turnovers per Game: []
```

```
Bonferroni Corrected p-value Threshold for all Statistics: 0.0017
```

```
Teams Which Meet Bonferroni Corrected Threshold
Field Goal %: []
3-Pt Field Goal %: []
Blocks per Game: []
Assists per Game: []
Steals per Game: []
Turnovers per Game: []
```

- After we used the Benjamini-Hochberg (B-H) procedure to adjust the threshold, no tests remained significant. Please note that throughout the project, we consider each team to be a single “test”. Bonferroni correction yielded similar results, as this threshold is guaranteed to be at least as low as with the B-H procedure.

- With respect to the individual tests, when compared to a naive threshold of 0.05, B-H and Bonferroni correction do not yield any significant tests at all. This means that when controlling for Family-Wise Error Rate and the False Discovery Rate, none of these tests are any longer statistically significant.
- In aggregate, this would mean that for no team in the entirety of the NBA are we able to determine that the presence of the star player is linked to an increase for any of the 6 performance statistics tested

Limitations, Further Testing, and p-hacking

- A major limitation which may have altered our results is the ability to run only 2000 repetitions of comparing shuffled data vs. regular data while performing A/B testing. Our computers were not powerful enough to run more repetitions without problems with memory or the computation taking an unreasonably long time.
- Given additional data and computing resources, I would extend the hypothesis testing for this question to include more than just the 2021-2022 seasons. Also, instead of just testing whether a "star player" (player on the team with highest PPG for season) creates various impacts on the game, I would like to analyze how starters for each team contribute to performance metrics related to their role; for example, I would like to analyze how starting point guards of different NBA teams contribute to the average ASTs per game, FG%, and 3PT%.
- The main way we avoided p-hacking was using threshold values other than alpha. Using BH and Bonferroni ensured that we were accounting for multiple tests, and keeping the chance of FDP/FWER at a limit.

Option C: Prediction with GLMs and nonparametric methods

Methods

What we're trying to predict

- We're trying to predict who won a game or not from a variety of performance statistics. For the random forest model, we have no choice over which features we're using given it's a nonparametric method. For the GLM's though we did go through a selection process in terms of which variables provided us with the best results. We ultimately selected points, defensive rebounds, assists, and blocks per game. We selected those by finding which features have a high correlation coefficient within our heatmap. Although there were features that had higher correlation than assists and blocks, those features would give us a singular matrix if used. These variables gave us the lowest AIC, likely because adding more stats didn't help as much as adding the extra variables brought the AIC down.

GLM

- We are ultimately selecting logistic regression because it fits what we want to measure the best. Here, we're going to use stats from a given game to predict whether a team won or lost the game. Logistic regression over the other GLMs handles binary data. We'll try the other models because they were in our hypothesis, but given that they don't handle binary, we'll likely get results over 1, which demonstrate they can't be used to predict wins or losses. We're assuming here that the wins and losses are binary data.
- We split up our data into two different situations for both our models, which is predicting based upon if they are an away team or if they are a home team. This is because we saw some differences within the correlation coefficients based upon this split, which goes along with the intuition that if you are the home team, you are more likely to win because of the familiarity of the environment.

Nonparametric

- We're choosing to use a random forest model for our nonparametric data. A random forest model suits our data structure given we have many types of stats, including binary ones, and we want to predict just wins and losses. Also, we can look if we want to and see how decisions are made after we use the model, which helps with interpretability over some other nonparametric methods.

Evaluation strategy

- For both of the models, we're going to use the accuracy and FDP. We think by using both of these we'll balance getting true positives and also not too many falsely identified wins.

- For the Logistic Model, we were interested in trying out two different models. Model One would be using the features that are most commonly used to evaluate how well a team is performing, which is by Points, Rebounds, and Assists. Model Two will be using the features that our heat correlation map showed us to be the most important features which are: Points, Defensive Rebounds, Assists, and Blocks. In order to select which model would be better to use we will be using the model selection metrics AIC and Deviance. We will be using AIC because this model selection is a good metric to use when there are more features in this case, as it penalizes models which use more features. In this case, model two has one more feature than model one. We will also be looking at the Deviance as in Model Two we are using Defensive Rebounds instead of overall rebounds which is used in Model one. This will help us see if the change in variables has helped increase the performance of the model, as deviance tells us how much worse the model is from a perfectly/saturated model.

Results

Random Forest:

- The Random Forest model provided an accuracy of approximately 80.65% for home teams and 81.5% for away teams. This suggests that the features used are relevant and have predictive power. However, there is room for improvement as the model incorrectly predicts about 20% of the data.

GLM (Negative Binomial & Poisson):

- Both the Negative Binomial and Poisson models seem to have significant coefficients, indicating that the features used have a statistically significant effect on the number of wins. The Negative Binomial model's use indicates over-dispersion in the count data, which it can handle better than the Poisson model.
- The Negative Binomial model showed a Log-Likelihood of -27748 and an AIC of 55503.84, while the Poisson model had a Log-Likelihood of-16087 and an AIC of 32182.69, suggesting that the Poisson model fits the data better due to a lower AIC.

GLM (Logistic Regression):

- The Logistic Regression model, with a Pseudo R-squared of .1858 and an AIC of **21391.834955171522**, indicates a decent model fit, although the R-squared value suggests that there is still unexplained variability in the outcomes.

Uncertainty Estimation

GLM Models:

- The standard errors for the coefficients provide an estimate of the uncertainty. For example, the coefficient for PTS_home in the logistic model is 1.0814 with a standard error of **0.033**, indicating that the true coefficient is most likely within the range of 1.016 to 1.147 with a 95% confidence interval.
- This uncertainty can be quantitatively stated as: "For every one unit increase in home points, the log odds of the home team winning increase by between 1.016 to 1.147, with a 95% level of confidence."

Discussion

Model Performance Comparison:

- Based on the AIC and Deviance between Model One and Model Two of the Logistic models, for both home and visitor teams, Model Two, with features of Points, Defensive Rebounds, Assists, and Blocks, performs much better.
- Based on AIC, the Logistic Regression model performed the best among the GLMs, suggesting it might be the most suitable for future predictions. We did not extensively test with NegBinomial and Poisson as we did with the logistic model, as we saw early on that these models perform poorly compared to Logistic. The other models gave us values of greater than 1, which doesn't make sense in the context of our model. We should be getting only values between 0 and 1. Only logistic really fits our data.
- The Random Forest Model has a higher accuracy of about 80% and lower FDP of about .27 for home teams and about 81% accuracy and .13 FDP for away teams. The Logistic model has an accuracy of 71% for both home and away and an FDP .37 for the away team and .33 for the home team. Therefore, the Random Forest Model seems to be the better model to use here, because it has higher accuracy rate and on average has a lower FDP. This may be because Random Forest reduces variance, increasing its predictive power and ability to model complex non-linear relationships.
- Confidence in applying these models to future datasets is moderate. The models showed reasonable predictive ability, but the accuracy indicates there's a significant portion of outcomes they cannot predict.

Model Fit:

- Both models fit the data reasonably well, but the fact that none of the models achieve extremely high accuracy or R-squared values indicates that there is complexity and variability in game outcomes that the models and features are not capturing.

Baysian/Frequentist:

- We used a frequentist approach on both of our models as we didn't use a prior. We didn't observe any difference besides the different accuracy in estimating wins.

Choosing Not to Provide Interpretations:

One may choose not to delve into interpretations if the primary goal of the modeling exercise is prediction rather than explanation. If the project's stakeholders are more interested in the accuracy of future game outcomes rather than understanding the underlying factors that drive those outcomes, the focus will be on the model's predictive performance. Another reason might be the lack of domain-specific knowledge to make accurate interpretations. Without deep expertise in basketball analytics, interpretations might be speculative and potentially misleading.

Limitations of the Models:

- The Random Forest model is a black-box and may not provide as much interpretability as GLMs. GLMs, while interpretable, make stronger assumptions about the data distribution and may not capture complex patterns as well as Random Forest.

Additional Data for Improvement:

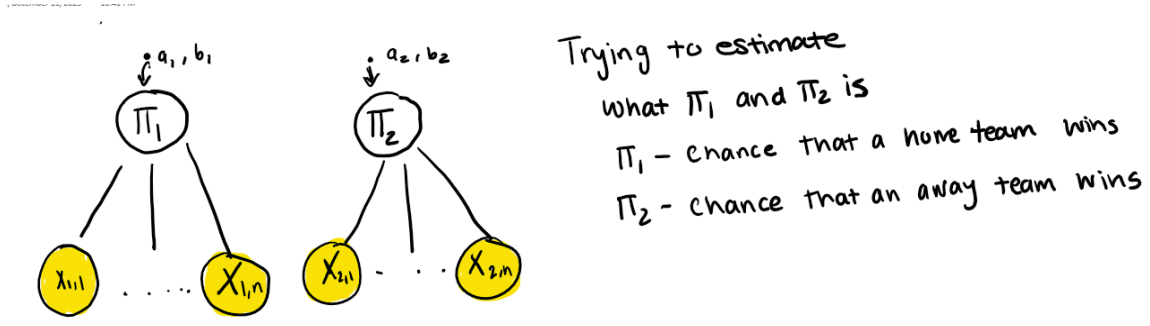
- Incorporating additional context-specific variables such as player injuries, rest days, home-court advantage, and individual player stats could potentially improve model performance.

Uncertainty in Results:

- The uncertainty in the results can be considered moderate given the accuracy levels achieved. Factors contributing to this uncertainty include potential data noise, overfitting, unaccounted variables, and inherent unpredictability in sports outcomes.

Option B: Bayesian Hierarchical Modeling

• Methods



We have fixed the parameters of the alpha and the beta from online research. The $X_{1,1}$ to $X_{1,n}$ are the observed result of if the home team wins and $X_{2,1}$ to $X_{2,n}$ are the observed result for each game that we have access to in the data frame. We want to estimate π_1 the win percentage of the

home team and π_2, the win percentage of the away team. This is over all the seasons we have data for given those variables and the data from our report.

We are not using a Bayesian mixture model.

We're picking a Beta prior for the win percentage because based on online research, we saw that if you are a home team then you have a 62.7% chance of winning, with a standard deviation of about 12.9. Those wins tend to taper out at the end. With this information and knowing π_1 and π_2 values should be bounded between 0 and 1, a beta distribution makes the most sense to use. That also makes sense with our domain knowledge of the area, given that NBA teams tend to win more at home than away.

We did online research to find estimates of the mean and standard deviation and then used the following equations to find the alpha and the beta, where m = mean and V = Variance:

$$alpha = ((m * m * (1 - m))/V) - m$$

$$beta = (((1 - m) * (1 - m) * m)/V) - (1 - m)$$

We ultimately get out alpha = 5.3 and out beta = 3.15 for home teams and alpha = 3.15 and beta = 5.3 for visitor teams.

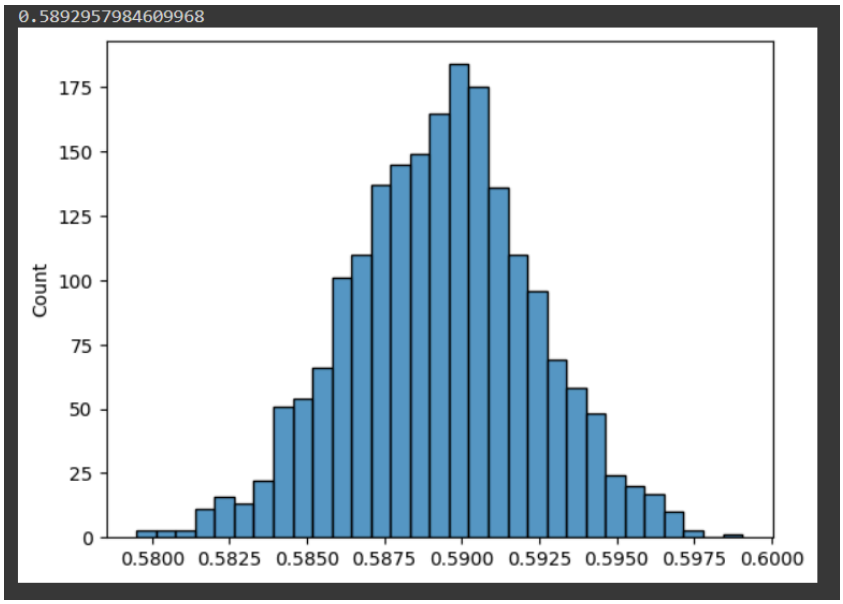
The mean we estimated from this link:

https://www.chicagobooth.edu/review/home-field-advantage-facts-and-fiction#:~:text=In%20basketball%2C%20NBA%20teams%20win,are%20 won%20by%20 home%20 teams.

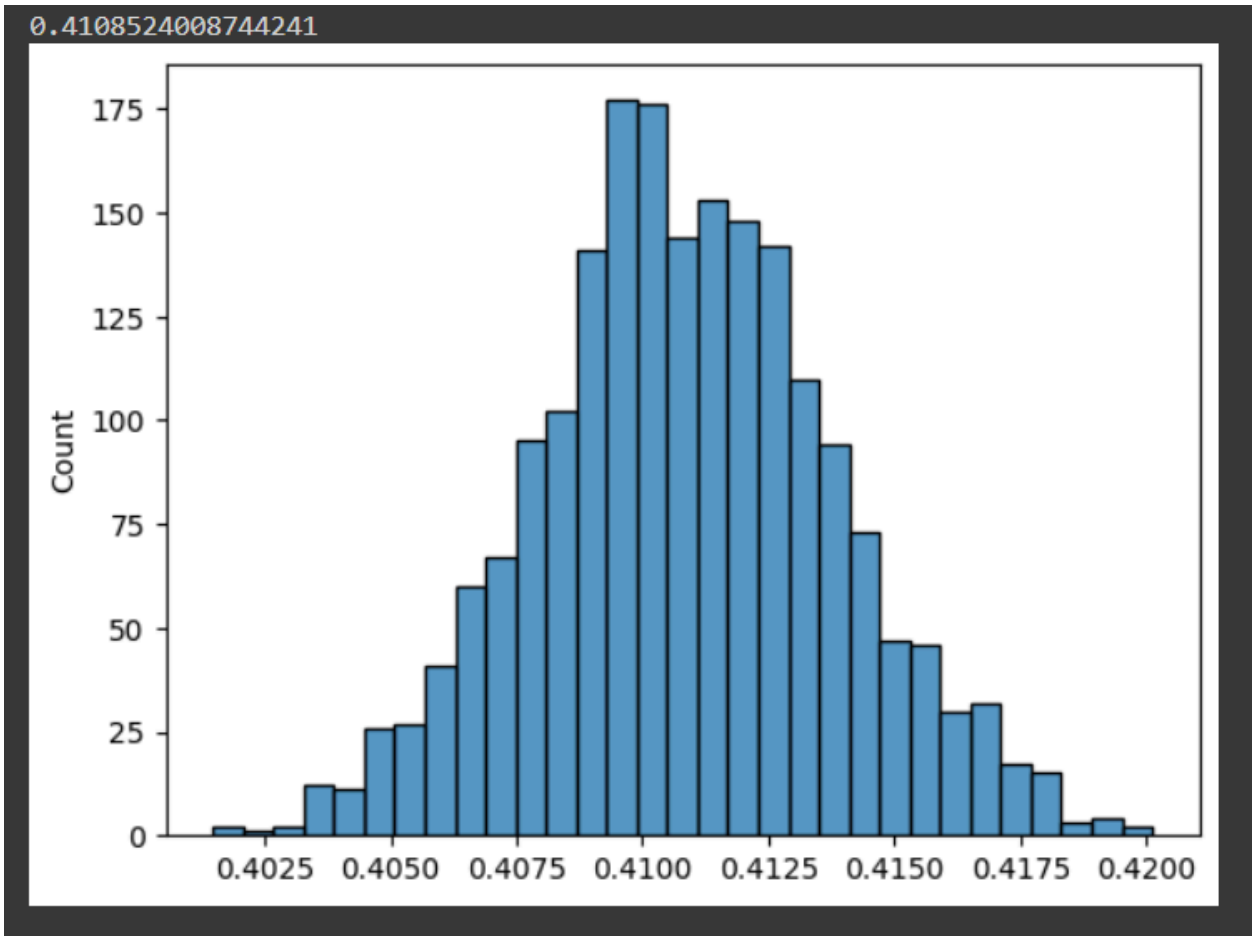
The standard deviation and number of games was from this link:
https://www.si.com/nba/2011/09/07/nba-parity

• Results

The results are actually very consistent with what we expected. We got an estimate of about 58% in terms of the chance of winning for a home team. That is actually a bit lower than the overall win estimate we found online for the long-term win percentage, so it's possible the traditional "home court advantage" in basketball has changed over time.



For away teams, we got a value of about 0.41. That means our model predicts that over time, we estimate that away teams have a 41% chance of winning. That's also consistent with the historical percentage we got from online research.



– Quantify the uncertainty in your estimates, and provide clear quantitative statements of the uncertainty in plain English.

• Discussion

Typically, the benefit from using Bayesian Hierarchical Modeling is the use of the priors to more accurately receive results. However, we are researching a fairly niche problem with very little common knowledge on the distributions of performance metrics.

We didn't have any trouble getting it to converge. We did have a bit of trouble finding the right prior for our model (originally thought we would just use uniform, but ended up with a beta distribution after some more online research.

We tried originally using a Bernoulli and then a Binomial distribution for the priors to feed into PyMC. Clearly that didn't work because we needed to input a beta distribution into the final Bernoulli model for us to use. We also just tried a uniform distribution, but ultimately chose to go with a normal prior because it provided us an opportunity to input the mean and standard

deviation that we found online. Putting those other values in allowed us to provide a model that likely does a better job of estimating the final percentage of wins for home/away teams.

It might be useful to try out even more priors. Within the prior, we might also have been able to come up with better values for the mean and standard deviation that were themselves distributions based on historical data over seasons. That might have provided us with even more accurate estimates than what we had with the normal distribution prior.

Conclusion:

First, to look at the models we created, we saw that it is possible to model to predict who is likely to win a game based on statistics and we saw which features are the best predictors. We were successfully able to use logistic regression as well as a random forest model to predict wins, with a fairly high accuracy for both and a low FDP. We were also able to narrow down the best statistics to use to predict a win. Second, we interestingly found that the star player didn't have a statistically significant effect on any of our tests across teams. That included FGP (field goal percentage), BLKs (blocks), TOs (turn overs), ASTs (number of assists), 3PT% (three-point field goal percentage), STLs (number of steals). Finally, with Bayesian hierarchical modeling, we concluded estimates for the total number of wins for home and away teams, based on historical data and the data we had access to.

Our findings are focused on one full basketball season. That means our models are applicable for that season. For the models, we did look at the data throughout the season (see our EDA section) and it seemed to be fairly consistent (didn't have obvious patterns that may have meant a certain part of the season needed a different model). So we're confident in applying our model to any point for that season. Our methods though are applicable to any season. As data starts to become public for a future season, we could easily fit our same models onto that data and use that new model for predictions in the next NBA season.

For the hypothesis testing, that data was also focused on just one season. So the conclusions on the significance of a "star player" are focused on that season. That analysis can be used for conclusions around how to form a team for the future season (if a new choice needs to be made around going with a different player lineup, paying the star player less, and others). Also, once future data is available, the same analysis can be easily replicated to be used for future seasons and subsequent decisions during a season (around benching a player for instance) or for future trades and star player decisions for seasons to come.

For the Bayesian hierarchical model, that data should be applicable to more. We used all the wins we had access to, which included many seasons as well as historical data to create the prior. That means that our estimate for the overall win percentage for home and away teams is useful to estimate over many seasons, which an average or just historical data wouldn't be able to.

The model predictor opens up a world of possible applications for using stats to predict a game win. Sports betting could be improved by using our model to predict a win, coaches could use the model to predict when they might win based on given stats and adjust accordingly, or fans could use it to predict when they might want to stay at a game or not based on the current stats.

For the hypothesis testing, interestingly, the "star player" might not be worth as much as coaches and fans seem to think. In terms of effect on various stats, the star player didn't have a statistically significant effect. In the future, teams may need to examine if what they pay their star players and if the number of minutes or attention on these players is warranted. This same analysis could be used to make decisions about whether a player should be played or not during a season, or if a line-up should be changed by copying our analysis.

For Bayesian, we now have an estimator for wins home/away based on historical and given data. That could be used by the NBA for predicting which teams (home or away) they should add more advertising money into (for instance advertise to home teams who are more likely to be winning and have more fans watching). It could also be used by teams to better predict home many home or away games they'll use, which could be useful info for them to determine performance.

We did not merge different data sources. We had plenty of data available in terms of different statistics for our teams available in the original data. If we want to expand our research questions in the future we may need to incorporate more.

Our main limitation was just computing power for hypothesis testing. In terms of the data, we might have added or at least analyzed more stats to use in our models (or that may have been used for our random forest model), but what we had gave us a fairly good prediction on test data as is.

It would be useful to use these models and conclusions in future seasons, and use real world feedback based on implementing these changes in iterating on the start we created. This same analysis could also be used in other sports, and might even be used in lower leagues if the same data is available. Lower children's and teen's leagues typically have no sense of using

statistics to identify where a player's weak points are. Teamwork can absolutely be encouraged at a younger age by identifying how certain players' presences on the court affects the team's performance. Overall, the use of these analysis tools in sports could help continue to shape the culture of sports around using data instead of instincts to drive team decisions.

We learned that many of the techniques used in class are useful with real world data as well. Using hypothesis testing with a better understanding of the impact of multiple hypothesis testing saved us from finding some results to be statistically significant when they shouldn't have been. Having no tests be statistically significant after controlling for FWER and FDR was certainly a jarring result. The models that we made we wouldn't have known to use an AIC to correct for overfitting without studying that in class. All of these tools were useful in shaping our understanding here.