Analysis of MLB All-Star Pitchers and the Impact of Elevation on Home Runs

Rodrigo Cochran, Noah Kaminer, Jonathan Wang

Data Overview

For the purpose of the following two research questions, two different sources of MLB data were used. The first research question employs a dataset called "Baseball Databank" which includes aggregated MLB data from 1871 to 2020. Moreover, the data included was collected via an observational study of the entire population of data from the specified years. This dataset is provided for free by Open Source Sports and downloaded via www.kaggle.com. While the downloaded dataset includes 24 individual datasets, ranging from pitching statistics to manager award data, only 2 of these datasets are utilized in the following analysis: Pitching.csv and AllStarFull.csv. Each row in the Pitching.csv represents a pitcher's pitching statistics in a given year, including whether they were selected as an All-Star that year.

The second research question utilizes a tool provided by www.baseballsavant.mlb.com, which allows the search of MLB.com's Statcast database. To retrieve the version of the dataset used in the subsequent analysis, the following filters were utilized:

Pitch Type = *{In play out(s), In play no out, In play run(s), In play pitchout runs(s)}*
Player Type = *Pitcher*
Group By *Player Type*
Season = *{2017, 2018, 2019, 2020, 2021}*
Sort By *Pitches*
Season Type = *Regular Season*
Sort Order = *Desc*

All other conditions remained with their default value. The retrieved dataset derives from altering the *Included Stats* to the following metrics: *events*: the outcome of the pitch, *stand*: whether the batter stands in the right or left-handed box, *p_throws*: the arm the pitcher uses to throw, *home_team*: the team name of the home team, *away_team*: the team name of the away team, *launch_speed*: the speed of the ball after contact with the bat, and *launch_angle*: the exit angle of the ball after contact with the bat. Once the data was filtered, it was downloaded via the "Download Data as Comma Separated Values File" button. The resulting dataset derives from an observational study of the entire population of MLB regular data from the 2017, 2018, 2019, 2020 or 2021 regular season. Each row represents a singular pitch that produced an in play event. The Statcast database we downloaded our data from only allowed us to pull 40,000 most recent pitches per year, so we use the 5 most recent seasons to get a csv with a total of 200,000 pitches.

The participants of both studies/collection of data–namely players and coaches–are fully aware of the collection of this data, but likely not in the aggregation of each particular dataset. It is also important to note that both final datasets do not introduce selection bias or convenience sampling due to the unabridged collection of MLB statistics. However, there may be some degree of measurement error in certain statistics, such as launch speed, due to the inherent inaccuracy of measurement tools, but these errors should be relatively constant across all of the data collected, so there is little room for concern.

Research Questions

Research Question 1: *Can one predict if a pitcher is chosen to be an All-Star in a given year based on pitching statistics?*

Currently, All-Star pitchers are selected by managers, whereas starting fielders are selected by fans, and reserves are selected by managers and players. This selection takes place via tallying of votes. Intuitively, the votes casted by managers to select All-Star pitchers likely relates to not only a pitcher's pitching statistics that season, but also a number of other factors, such as popularity and league dynamics.

The research question above seeks to not only explore the current objectivity in determining All-Star pitchers, but also the feasibility of replacing the voting system with a more objective, statistics-based classification of All-Star pitchers. Prediction with generalized linear models (GLMs) and nonparametric models is a natural fit for this type of research question because it signifies classification and prediction of a binary variable based on continuous variables. Specifically, the nonparametric used will be Random Forest because of the ability to split the data (pitchers) based on numerous pitching statistic thresholds. The GLM used will be Logistic Regression because being an All-Star in a given year is a binary categorical variable. Both of these models will result in a prediction of whether a pitcher is an All-Star in a given year.

Research Question 2: *What is the causal effect of stadium elevation, specifically playing at the Colorado Rockies stadium, on the number of Home Runs scored?*

There are 30 MLB baseball stadiums currently in use. These stadiums can be drastically different from each other in terms of size, elevation, and weather conditions. This question seeks to analyze the effect of elevation on the number of Home Runs at a stadium. Specifically, sitting at 5,211 feet above sea level, the Colorado Rockies stadium (Coors Field) is approximately 4,100 feet further above sea level than any other stadium in the MLB (Arizona Diamondback stadium is 1,059 feet above sea level). While it is assumed that baseballs fly further in higher altitudes, this question seeks to quantify the impact. Causal inference works perfectly for this question because pitch-by-pitch data can be used to model the causal effect of altitude, while also accounting for confounding variables such as statistics relating to the time of impact of each pitch resulting in an in play event. This analysis seeks to help value both hitters and pitchers for the Colorado Rockies. Specifically, if the altitude of the Colorado Rockies stadium has a large causal effect on the number of Home Runs, then their hitters may be overvalued and their pitchers undervalued. The following strategies will help to quantify these impacts and normalize the Rockies' players against other players in the league.

Exploratory Data Analysis (EDA)

Research Question 1

Despite some of the data in the dataset for research question 1 dating back to 1871, a number of factors, including but not limited to the changing of team names and rules, encouraged utilizing data beginning in 1985. Moreover, the 1994 lockout caused by negotiations and the curtailed 2020 season caused by COVID-19 motivated removing these seasons from the following data analysis. Thus, the final dataset only includes data from 1985 - 1993 and from 1995 - 2019, however, it is expected that this modification will have little-to-no impact on the results of the model and the inferences made thereafter. The Pitching.csv and AllstarFull.csv datasets were merged on both *yearID* and *playerID*. A number of columns were then removed due to irrelevance to the question at hand; the resulting dataset included the following columns in the dataset: *playerID, yearID, teamID, SHO, H, ER, HR, BB, SO, BAOpp, ERA, WP, HBP, R, GIDP, is_all_star*. All rows that contained null values were dropped because in comparison to the total number of rows the number of rows that contained null values was small; the number of rows in the dataset decreased from 22,153 to 22,129. We decided to *keep playerID, yearID*, and *teamID* in the initial data cleaning to keep track of the players that would be pulled from the All-Star database, but these variables were not used in the training of the model.

We looked at a correlation chart of all of the variables in our dataset to see which variables were impacting each other, and their resulting relationships (Figure 1.1). Furthermore, to see the resulting relationships to the output variable, *is_all_star*, we created a side-by-side boxplot that examined the distribution–specifically focusing attention on the median–of each variable for both all-stars and non-all-stars (Figure 1.2).

Research Question 2

In order to prepare the data, we selected the columns *events, stand, p_throws, home_team, plate_x, plate_z, launch_speed, launch_angle, and effective_speed.* We used the *home_team* column to create the binary treatment variable that was equal to 1 if the game was played at Coors Field and 0 if else. We also used *events* to create our binary outcome variable *HR* that was 1 if the pitch resulted in a home run and 0 if else. Next, we broke up the data into pitches thrown by right and left-handers using *p_throws* and dropped the *events, home_team, and p_throws* columns. Finally, we one hot encoded the *stand* column to indicate whether the batter was in the right or left-handed batters box.

We decided to first determine if the fact that higher elevation causes baseballs to fly further actually translates to more home runs. To do this, we created a bar graph displaying the number of home runs at the Colorado Rockies' stadium compared to other stadiums (Figure 2.1). Note that homeruns are separated by pitching arm to potentially aid any other questions that we may have in subsequent analyses. We also knew that baseballs flying further in high altitudes would correlate with launch speeds and launch angles. Thus, we compared the Rockies' stadium to other stadiums on both of these statistics (Figure 2.2) via a boxplot to better understand this correlation, if one, in fact, exists.

Inference and Decisions

Research Question 1

Methods

For our first research question, we are trying to predict if a pitcher is going to be an All-Star in a given season, based on their pitching metrics. For our features, we have selected the following relevant pitching statistics because of their availability and comprehensiveness of a pitcher's ability. The features chosen are: *SHO* (# shutouts), *H* (# hits against pitcher), *ER* (# earned runs for a pitcher), *HR* (# home runs against this pitcher), *BB* (# walks), *SO* (# strikeouts), *BAOpp* (opponent's batting average), *ERA* (earned runs average), *WP* (# wild pitches), *HBP* (# beanballs), *R* (# runs given up), and *GIDP* (# groundouts that were double plays). To accurately model the relationships between our features and output, we are using a nonparametric model and a GLM. The nonparametric model used is **Random Forest** because it is a large dataset with an unknown underlying distribution and Random Forest makes predictions via decisions on numerous variable thresholds. The GLM will be **Logistic Regression** because being an All-Star in a given year is a binary categorical variable, and logistic regression is a powerful model for this type of prediction. To ensure validity, each model will be evaluated on 3 metrics: testing accuracy, precision, and recall. For each of these metrics, a higher value is indicative of a more successful model.

Results

For the **Random Forest** pre-optimization, we achieved the following validity metrics: a testing accuracy of **0.8348**, a precision of **0.8174**, and a recall of **0.8755**. Our interpretation of this is that with default hyperparameters, the model performed fairly well with accurately predicting All-Star labels. The high accuracy is accompanied with high recall and precision, meaning that it returns more relevant results than non-relevant results and that most of the relevant results are returned. This is indicative of high-quality results.

For the **Logistic Regression** pre-optimization, we achieved the following validity metrics: a testing accuracy of **0.8486**, a precision of **0.8354**, and a recall of **0.8800**. Our interpretation of this is very similar to the Random Forest. The Logistic Regression model performed well with accurately predicting All-Star labels. The **Logistic Regression** maintained high precision and recall, meaning that it returns more relevant results than irrelevant ones and that most of the relevant results are returned. This is indicative of high-quality results.

After evaluating the results of the two models, we decided that even though our results were acceptable, we should run an optimization, using **Random Search** to see if we can achieve even better results. We utilized a similar implementation to that of Will Koehrsen, from the article *Hyperparameter Tuning the Random Forest in Python* (Towards Data Science) for both the **Random Forest** model and the **Logistic Regression** model.

For the **Random Forest** post-optimization, we achieved the following validity metrics: a testing accuracy of **0.8394**, a precision of **0.8189**, and a recall of **0.8844**. Our interpretation of this is that the optimized model returned very similar results to the initial model, but each of the 3 metrics did, in fact, increase, so we can say the optimization was successful.

For the **Logistic Regression** post-optimization, we achieved the following validity metrics: a testing accuracy of **0.8486**, a precision of **0.8354**, and a recall of **0.8800**. Our interpretation of this result is that the optimized model returned nearly identical results to the initial model, but each of the 3 metrics did, in fact, marginally increase, so we can say the optimization was successful.

To further evaluate the GLM predictions, we looked at the quantitative uncertainty of our model. By taking samples of the data and running the tuned **Logistic Regression** model on each of these, we can create a confidence interval that represents the uncertainty of our GLM. After taking 100 samples that were each trained on **80%** of the data, we obtained an interval that ranged from a minimum of approximately **82.23%** test accuracy to a maximum of approximately **90.83%** test accuracy, with a mean of approximately **85.68%.** This means that, at the first percentile, we are predicting All-Star status with an **82%** test accuracy score, and marginally higher value of **82.28%** at the fifth percentile, but while we can expect results to have an approximately **85%** test accuracy for any sample, we are **95%** confident that the test accuracy would lie on the interval **[82.28%, 89.08%]**.

Discussion

The two models performed nearly identically, but the **Logistic Regression** model had a marginally higher accuracy than the **Random Forest** model; they differed only by a fraction of a percent. We are confident in applying this to future datasets because with cross-validation and hyperparameter tuning, we can achieve a high accuracy without overfitting.

In each model we saw many similarities: nearly identical testing accuracy, precision, and recall. We find this to be interesting as both models perform exceptionally well, but for different reasons. Given the ‘black box’ properties that come with the **Random Forest**, it is difficult to understand why they perform so similarly, but the similarities may be due to the actual relationship of the variables in the data.

Even though we achieved fairly good results, there were some limitations to each model. One of the limitations of **Random Forest** is the fact that it is a ‘black box’ algorithm, so we don’t necessarily know how it’s creating its decisions, but this limitation would not be incredibly limiting as we are able to perform cross validation and other tuning on the model. The main limitation of the **Logistic Regression** model is that it assumes linearity and no multicollinearity between independent variables. It can underperform on data with variables that do not meet these requirements. In our model, the exact relationship between the independent and dependent variables is unknown, however, this did not appear to limit the effectiveness of the model; we can assume that this limitation would not be significant in our tests on future data.

To combat this, we think there are two variables that might help improve each model. These two variables would be if the player was an All-Star in the previous year, and a valuation metric that would monetarily value each player. Obviously, the most valuable players have the highest probability of being selected as an All-Star, so this variable could help a lot. Additionally, the fact of being an All-Star in the previous year may reveal more about the value of the player, which could help with predicting their All-Star contendabilty for the following year.

Research Question 2

Methods

We decided to break up the treatment groups into balls hit into play at the Colorado Rockies stadium, Coors Field and balls hit into play at other MLB stadiums. Our outcome variable was a binary one that equaled 1 if the ball hit into play ended up becoming a Homerun.

The confounding variables include anything that impacts how the bat comes off of the ball. We included the batter's stance, the horizontal position of the pitch at home plate, the vertical position of the pitch at home plate, the launch speed of the ball off of the bat, the launch angle after contact, and the effective speed as confounders. These known confounding variables are the most impactful factors that influence whether the batted ball will end up being a homerun. Besides that, air density is one of the only other variables that impacts the chance the ball is a home run, which is accounted for in our treatment variable. We can assume unconfoundedness because given these confounders, home runs and whether the ball is batted at Coors Field are conditionally independent.

First, we broke up the dataset into pitches that were thrown by Right-Handed Pitchers and those that were thrown by Left-Handed Pitchers. This was to adjust for the differences in movement that pitches thrown from different sides have. In order to adjust for our confounding variables, we initially decided to use the matching technique. Since some of our variables like launch speed and launch angle are continuous variables, we discretized them into four categories based on which quartile they fell into. After, we were left with only categorical and binary variables. Once our tables for right and left-handed pitchers were prepared, we ran a matching algorithm to find similar pitches thrown at Coors Field with ones thrown at other stadiums. After running several variations of the algorithm, we realized that it was too computationally expensive and would take half a day to run. Since we did not want to make our data too sparse, we decided to take a different approach and opted to estimate the average treatment effect through inverse propensity weighting. We used logistic regression to obtain the propensity scores and then used the IPW formula to get the estimate. We did not have any colliders in our data.

Results

The IPW estimate of the average treatment effect of pitching at Coors Field was 1.571 for Right-Handed Pitchers and 2.185 for Left-Handed Pitchers. Since the treatment effect for both right and left handers was positive, we can assume that batting at Coors Field leads to more home runs. More specifically, if we assume that home runs are unconfounded given our confounders, then the IPW estimate of playing at Coors Field leads to 1.571 more home runs given up by Right-Handed Pitchers and 2.185 more home runs given up by Left-Handed Pitchers. These estimates are also under the assumption that wind or temperature does not play a significant role in homeruns.

Some uncertainty in our estimates comes from the skill level of the players on the Colorado Rockies roster over the past 5 years. In 2017 and 2018 they had a positive record, but have had a losing record for the past 3 seasons. If Rockies batters are worse than league average, then home runs could be underestimated because they hit fewer and make up most of the treatment data set. If Rockies pitchers are worse, then home runs could be overestimated for similar reasons. The former would lead to a lower observable causal effect than in reality, and the latter would lead to a higher causal effect.

Discussion

Some of the limitations in our methods were in our inability to implement our matching algorithm. We couldn't find an efficient way to sort through all of the rows in our set and match, as there was a lot of data. We also encountered limitations when trying to pull data from Statcast because we could only get the most recent 40,000 pitches per year for each query. Other limitations include not being able to quantify the ability of the pitcher who gave up the home run and the batter who hit it.

Additional data on weather would be useful for answering this question because factors like wind speed and wind direction can heavily influence how far a ball travels in the air. Weather conditions would also be useful because things like rain also add extra resistance to a ball's path of movement. Temperature could also be a useful variable because a baseball would travel further in warm air because there air density is lower. All of these factors would be useful because they could potentially act as confounders in our data.

We are confident that there is a causal relationship between Coors Field leading to more home runs. It is supported by our model and makes sense because its high altitude means that there is lower air density, which would allow a batted ball to carry further.

Conclusions

Our findings from research question 1 illustrate that pitching statistics are very good predictors of whether a pitcher will be an All-Star in a given year. The findings from research question 2 suggest that there is, in fact, an impact on the number of home runs due to the elevation of Coors Field, as compared to other fields in the MLB. These results are not only what we predicted while developing our research questions, but also significant in the impact they could have on the future of the MLB.

It is first important to note that our findings are considered generalizable. While cleaning the data for question 1, it was necessary to use a random sample of all of the data for more accessible data and results. Nonetheless, randomization ensures that the model can accurately predict any wealth of new data. During the collection of data for the second question, we were limited to the last 40,000 pitches of each season due to downloading capabilities. While this data only encompassed about the last half of the season, we assumed a constant distribution of pitches and home runs throughout the season. Combined, we believe that the methods used, make the aforementioned results applicable to any new MLB data.

Unfortunately, there were a couple of factors that we could not account for in our analysis. For question 1, more data points would have been useful, but the changing of rules, along with other factors, made it difficult to go back any further without complications. Moreover, it would have been very useful to use every pitch in the last five regular seasons, as opposed to only the last 40,000 pitches of each season because it would have made our model account for pitches in different seasons and weather conditions throughout the year.

While the two research questions both involve baseball statistics and predictions, they involve data with significantly different granularity, and thus we refrained from merging the two datasets. This created a rather separate approach to the questions and implications of the findings, but nonetheless, sought to help us better understand if baseball is really a game of numbers, as the movie *Moneyball* suggests.

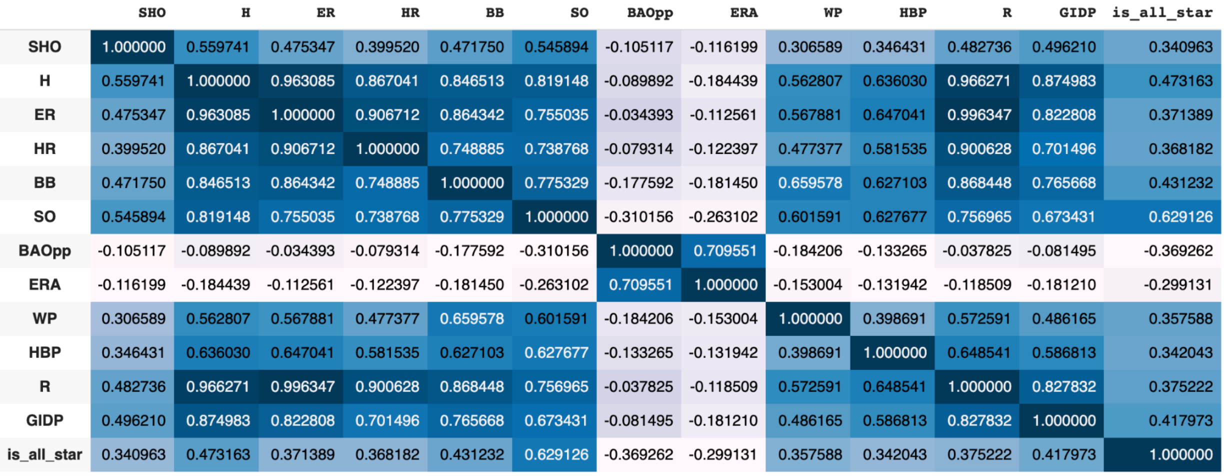
The results of question 1 indicate that the voting technique used to determine All-Star pitchers may, in fact, be outdated. Considering we were able to very accurately predict All-Star pitchers based on just a few pitching statistics, we wonder if the MLB should opt to purely use statistics instead of voting. While this seems to be a reasonable conclusion to our results, it is important to understand the consequences that a drastic change like this might have on the league and league-dynamics. Specifically, as a follow-up to these results, it would be interesting to study the impacts that changing the method of choosing All-Stars would have on fan involvement in the MLB. Would the switch to an algorithm make baseball fans across the country become less involved in the game?

The results from question 2 possibly suggest a less drastic and more reasonable call to action. That is, since the results suggest it is easier to hit a home run at Coors Field, then maybe the MLB should consider extending the walls of the field to make the chance of hitting long balls more in line with other fields. Another possible approach could be to have baseball managers account for this when evaluating the Rockies pitchers and batters, as this could have significant implications for Rockies players' trade value. For example, Nolan Arenado was a five time all-star third baseman while on the Colorado Rockies before he was traded to the St. Louis Cardinals. During those five high-performance seasons with the Rockies, he hit 37 to 42 home runs each season. In 2021 with the Cardinals, he only hit 34 home runs and his other hitting stats declined significantly. Even though this may not be purely because he is playing at a lower altitude, it is still an interesting case study into the effect that playing at Coors Field can have on players.

While we believe that the results in this report certainly have huge implications for baseball fans and the MLB, we know that research into this field is far from complete. Specifically, given the time and computing capability, we believe that implementing a proper matching algorithm to research question 2 would provide even more interesting and accurate results. Accordingly, we believe that it would be interesting to change the outcome variable from home runs to a more complex statistic like slugging or wOBA, which are more granular statistics than home runs because they account for the number of bases received from a given pitch, rather than just home runs.

Appendix

Figure 1.1



	SHO	H	ER	HR	BB	SO	BAOpp	ERA	WP	HBP	R	GIDP	is_all_star
SHO	1.000000	0.559741	0.475347	0.399520	0.471750	0.545894	-0.105117	-0.116199	0.306589	0.346431	0.482736	0.496210	0.340963
H	0.559741	1.000000	0.963085	0.867041	0.846513	0.819148	-0.089892	-0.184439	0.562807	0.636030	0.966271	0.874983	0.473163
ER	0.475347	0.963085	1.000000	0.906712	0.864342	0.755035	-0.034393	-0.112561	0.567881	0.647041	0.996347	0.822808	0.371389
HR	0.399520	0.867041	0.906712	1.000000	0.748885	0.738768	-0.079314	-0.122397	0.477377	0.581535	0.900628	0.701496	0.368182
BB	0.471750	0.846513	0.864342	0.748885	1.000000	0.775329	-0.177592	-0.181450	0.659578	0.627103	0.868448	0.765668	0.431232
SO	0.545894	0.819148	0.755035	0.738768	0.775329	1.000000	-0.310156	-0.263102	0.601591	0.627677	0.756965	0.673431	0.629126
BAOpp	-0.105117	-0.089892	-0.034393	-0.079314	-0.177592	-0.310156	1.000000	0.709551	-0.184206	-0.133265	-0.037825	-0.081495	-0.369262
ERA	-0.116199	-0.184439	-0.112561	-0.122397	-0.181450	-0.263102	0.709551	1.000000	-0.153004	-0.131942	-0.118509	-0.181210	-0.299131
WP	0.306589	0.562807	0.567881	0.477377	0.659578	0.601591	-0.184206	-0.153004	1.000000	0.398691	0.572591	0.486165	0.357588
HBP	0.346431	0.636030	0.647041	0.581535	0.627103	0.627677	-0.133265	-0.131942	0.398691	1.000000	0.648541	0.586813	0.342043
R	0.482736	0.966271	0.996347	0.900628	0.868448	0.756965	-0.037825	-0.118509	0.572591	0.648541	1.000000	0.827832	0.375222
GIDP	0.496210	0.874983	0.822808	0.701496	0.765668	0.673431	-0.081495	-0.181210	0.486165	0.586813	0.827832	1.000000	0.417973
is_all_star	0.340963	0.473163	0.371389	0.368182	0.431232	0.629126	-0.369262	-0.299131	0.357588	0.342043	0.375222	0.417973	1.000000

Figure 1.2

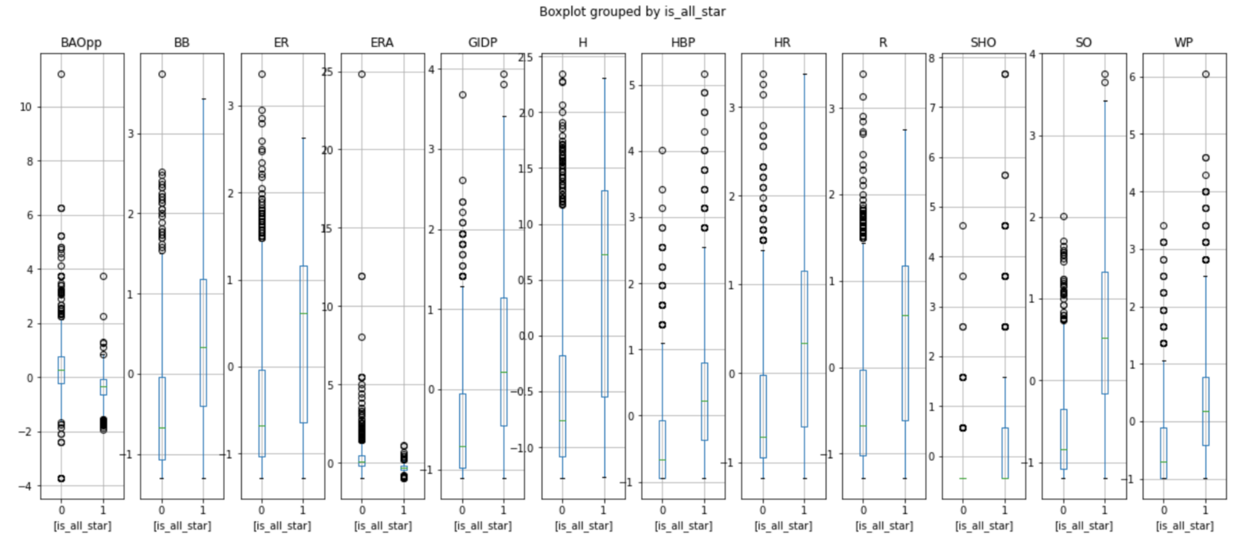


Figure 2.1

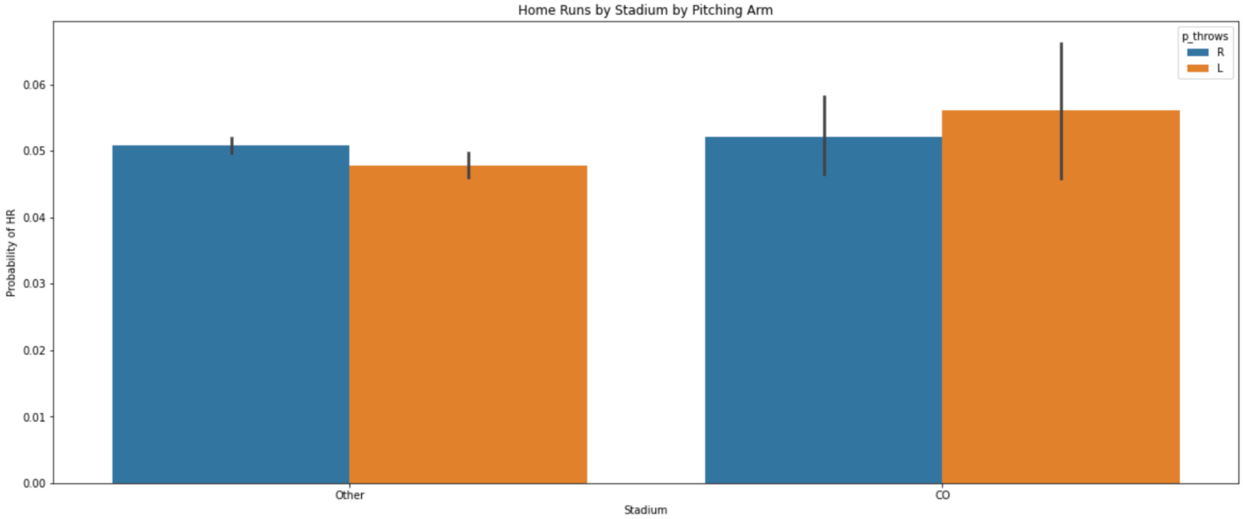
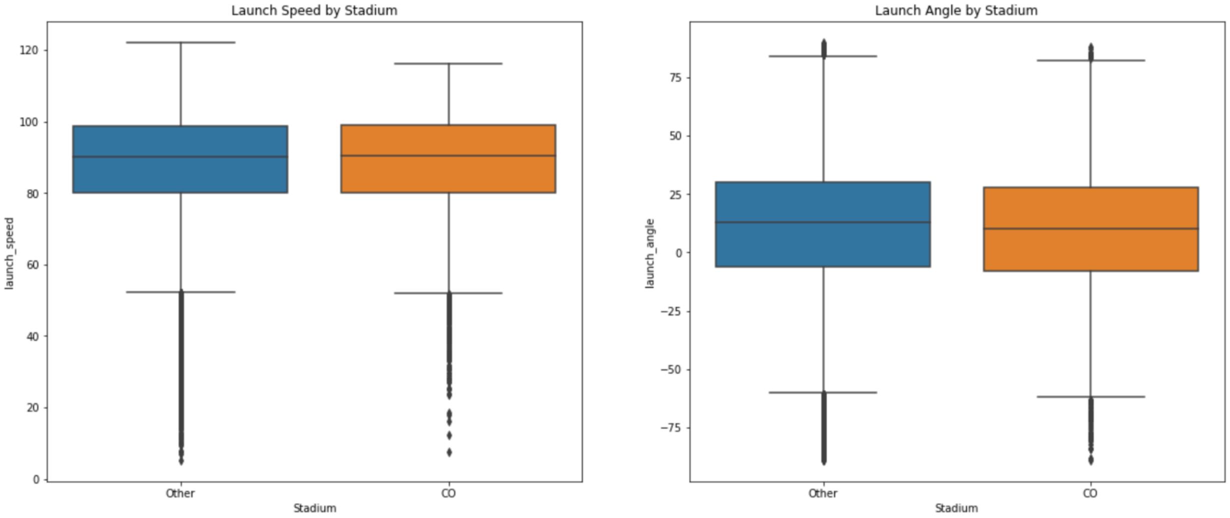

Home Runs by Stadium by Pitching Arm
p_throws
R
L
Probability of HR
0.06
0.05
0.04
0.03
0.02
0.01
0.00
Other
CO
Stadium

Figure 2.2


Launch Speed by Stadium
launch_speed
120
100
80
60
40
20
0
Other
CO
Stadium
Launch Angle by Stadium
launch_angle
75
50
25
0
−25
−50
−75
Other
CO
Stadium

Data 102 Project Checkpoint 1

Data Cleaning:

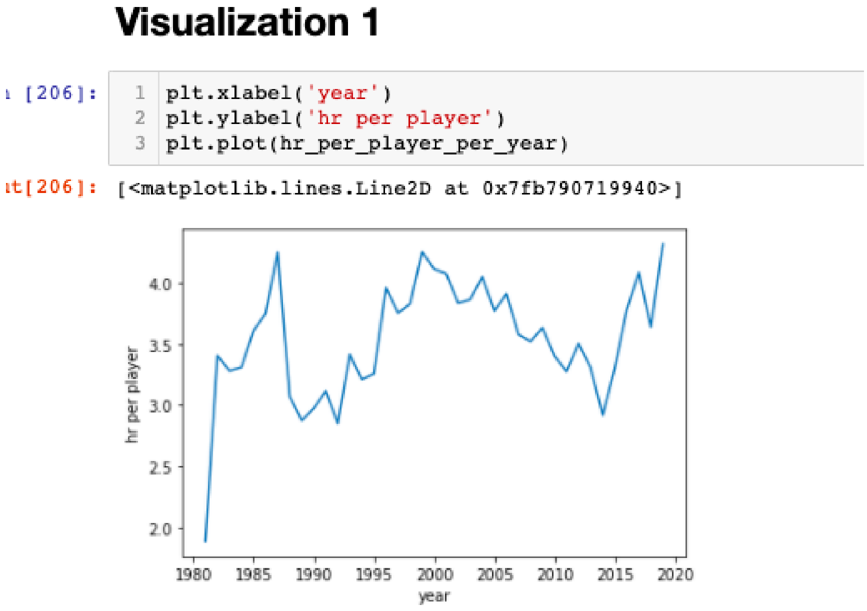
After looking through the various datasets in question, it was determined that each dataset would be limited to data collected from 1985 to 2016. This is a necessary step in our analysis because before 1985 there were many incongruencies in the data that would make it hard to interpret. Also, for the second question, the data analysis requires the Salaries dataset, which was only collected in the aforementioned dates. We do not believe that this will have any negative impacts on the model or the inferences made from the model because the data used will still include 31 years of data, which is plenty for this type of analysis.

Due to the complexity and breadth of this entire dataset, the data comes in numerous CSVs with variables and data that will be unimportant in the complete analysis provided in the next steps. Moreover, this analysis will utilize the following CSVs from the complete dataset: People, AllstarFull, Salaries, Batting, Pitching, and Teams. Within each of these tables, variables have been omitted. The aforementioned tables have been merged for a more complete analysis of the two questions at hand.

EDA:

The first two visualizations were made in hopes of answering the question “do hitting statistics cause higher salary?”, but in this process we found that there may be a more plausible question to answer and that would be “does paying the players more affect the outcome of their world series race?”. This EDA thus changed our first question to the one that follows.

What is the causal effect of a team’s average salary on its ability to win or probability of a world series?

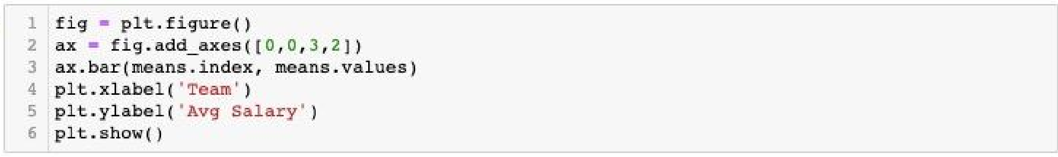


Visualization 1

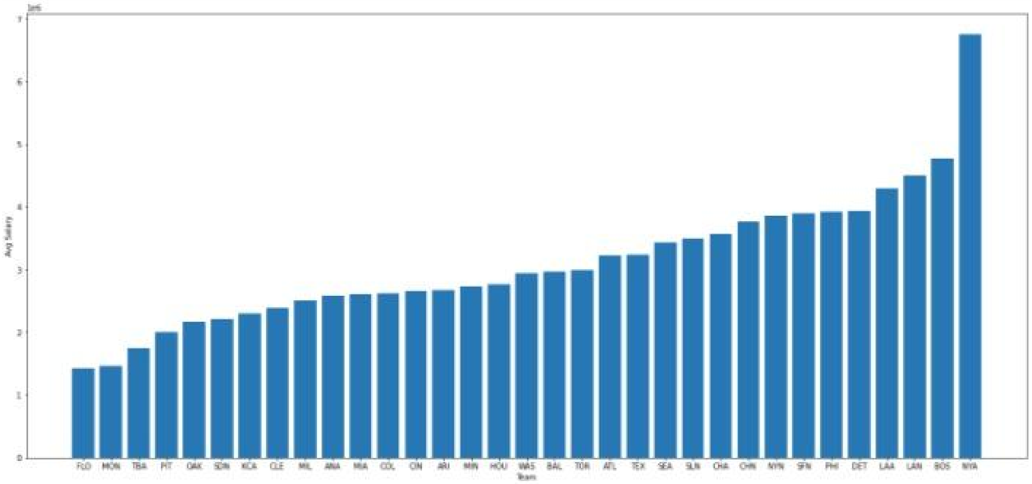
```
ı [206]:  plt.xlabel('year')
          plt.ylabel('hr per player')
          plt.plot(hr_per_player_per_year)

ıt[206]:  [<matplotlib.lines.Line2D at 0x7fb790719940>]
```

The visualization above displays homeruns per player over time. As shown, there are clearly significant changes in homerun statistics over the last 30+ years. Notably, there was a drastic increase around 1984-1987 and then a subsequent decrease around 1988. Another notable trend is that from around 2000 to 2014, the number of homeruns per player decreases, but in 2015 there happens to be a large, sudden increase. While we are currently unsure of the exact reasons of trends, we conjecture that these fluctuations are most likely a result of various historical events, including but not limited to wars, policy changes and rule changes, but all together suggest that history, particularly certain historical events, plays an important role in the homeruns per player statistic and the outcome of a baseball game. From this, we have concluded that time may very well be a confounding variable in our analysis and certainly something that we need to account for moving forward. Understanding this confounder makes us believe that there are variables that play an important role in the statistics that we will later perform our analysis on, and thus, we will seek to find more of them as well. This will certainly continue to shape how we approach this question and determine if we can, in fact, account for all of the confounding variables or if we need to reevaluate our approach.

Visualization 2



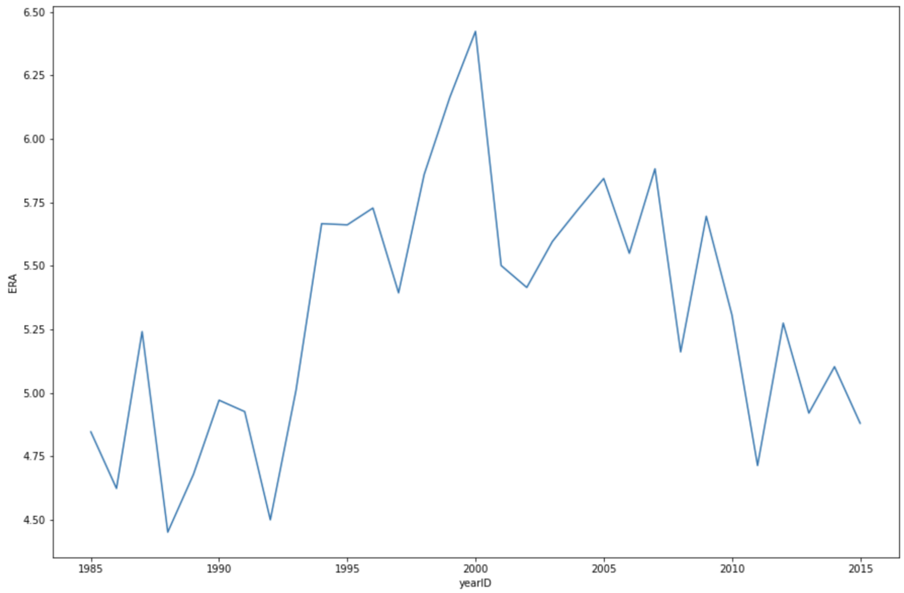
```
fig = plt.figure()
ax = fig.add_axes([0,0,3,2])
ax.bar(means.index, means.values)
plt.xlabel('Team')
plt.ylabel('Avg Salary')
plt.show()
```



The above visualization displays every team’s average salary over the given time period. Note that there are 35 teams listed, despite only 30 teams currently being in the league. This is due to teams entering and exiting, as well as teams changing names. We stratified by year 2000 or later because the turn of the century marked a historical moment that excluded a lockout year such as 1994, or even earlier ones such as 1985 where data was much different than usual years. This visualization shows which teams have historically spent the most money on their players, and we hope to use this as a confirmation of whether salary is deemed by the skill level or if the skill level could be determined from salary. As shown, the team labeled *NYA*, which represents the New York Yankees, has the highest average salary by an immense amount. This has long been known by baseball fans and players across the country. However, aside from the Yankees, there does not seem to be a drastic difference between any of the other teams, except from the highest to lowest salaried teams. Another interesting take from this graph that certainly lends itself to our question is the fact that the Yankees, Red Sox, and Dodgers are the teams with the top 3 average salaries, and are also 1, 3, and 6, respectively in World Series wins.

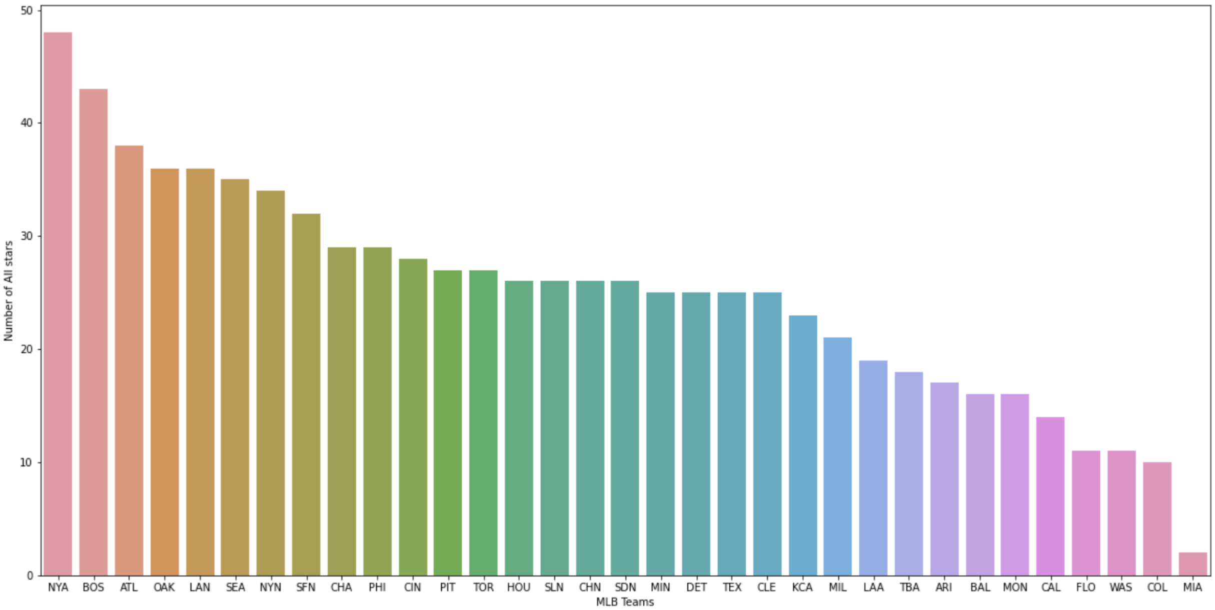
While the visualization shows that teams that have had more money seem to be the historically 'best' teams, we're interested in determining if money/salary has a causal effect on winning world series', a question which will certainly require a deeper analysis into the various aforementioned variables, including confounding variables.

Can we predict if a pitcher is going to be an all star in a given year based on pitching statistics? *(Changing to GLM/Nonparametric from Non-binary decision making based on proposal feedback)



This line plot shows earned run averages (ERA) over time. An earned run is defined as any player that scores as a result of a pitcher letting them get on base from a walk or hit. It does not factor runs that got on base because of a fielding error or passed ball and is a standard statistic used to evaluate a pitcher's performance. It is adjusted per nine innings so that starting pitchers who pitch 4-7 innings can be compared with relievers or closers who only pitch 1-3 innings. As shown in the graph, ERA starts off relatively low in the late 80's and early 90’s only for it to have a sudden spike in the late 90’s and early 2000’s. This is likely due to the increased use of steroids from batters across the MLB which increased their ability to hit harder and score more runs. After steroids are cracked down on in the league, earned runs steadily decrease.

This plot is relevant to our second research question about how pitching statistics affect all star appearances because we may have to look at individual seasons as opposed to a 20-year span like the graph above. The large difference in ERA between the 2000’s and late 80’s will make analysis challenging, as it could create limitations in what our nonparametric model could predict. In addition, these effects from steroid usage are likely present in other pitching statistics like strikeouts, walks, and saves.



The second plot details the number of pitchers that were selected as all stars from each MLB team from 1985 until 2016. Large market teams like the Yakees (NYA), Red Sox (BOS), and Dodgers (LAN) have some of the most successful pitchers over this time interval. Teams that were established later that 1985 like the Colorado Rockies (COL) and Miami Marlins (MIA) have had less time to produce all star talent and have smaller fan bases.

This discrepancy in all star selections across teams may play a significant role in our analysis trying to answer our second research question. Since all star selections are decided based on an unknown combination of player, manager, and fan influence, teams with larger fan bases and more media popularity may receive an unfair advantage. This could result in a pitcher with worse pitching statistics being chosen because they play for a team with a larger fan base and thus, more available votes in the fan polls. Since our question is just striving to use purely pitching statistics to predict all star appearances, some adjustments may need to be made later to control for which team each pitcher plays for.

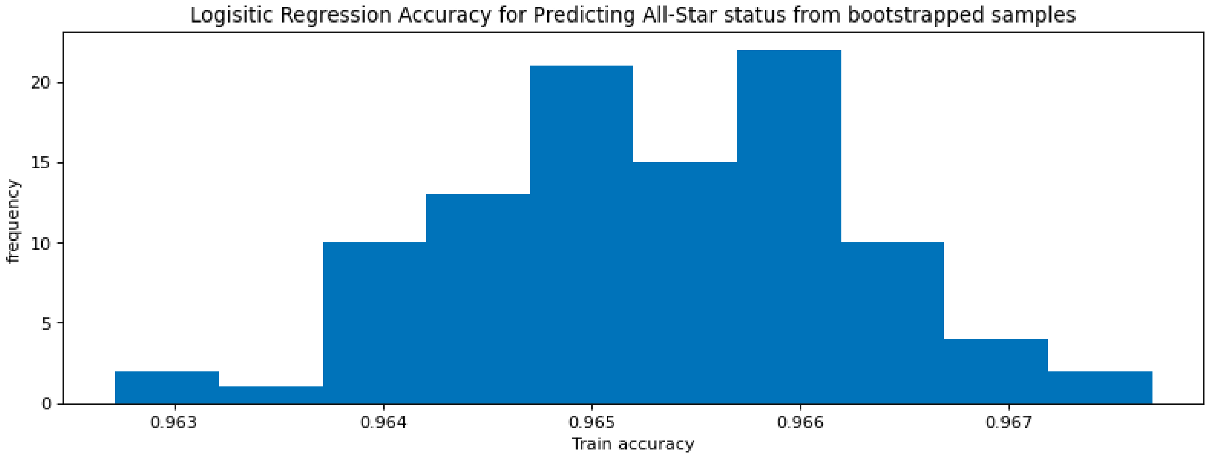
Data C102 Checkpoint 2

Can we predict if a pitcher is going to be an All Star in a given year based on pitching statistics?

In order to accurately research the question above, a GLM and nonparametric method was used to predict whether a pitcher is going to be an All Star. Since being an All star is a binary variable–a pitcher is either an All Star or not in a given year–the GLM utilized was Logistic Regression and the Nonparametric model used was Random Forests. After splitting the data into train and test sets, the 'LogisticRegression' and 'RandomForestClassifier' models from the scikit-learn package were trained on the training data. The Logistic Regression model produced a training data accuracy of 96.63% and a testing data accuracy of 96.29%. The Random Forests Classifier produced a training data accuracy of 99.99% and a testing accuracy of 96.58%. This means that the Logistic Regression model accurately predicted whether a pitcher would be an All Star 96.29% of the time. Accordingly, the Random Forest Classification model accurately predicted whether a pitcher would be an All Star 96.58% of the time. These results are incredibly good and illustrate that, while being an All Star is determined by votes from coaches, fans, and players, ultimately pitching statistics determines those votes. In other words, pitching statistics, as defined in the EDA section, play a large role when coaches, fans, and players cast their votes for All Star pitchers.

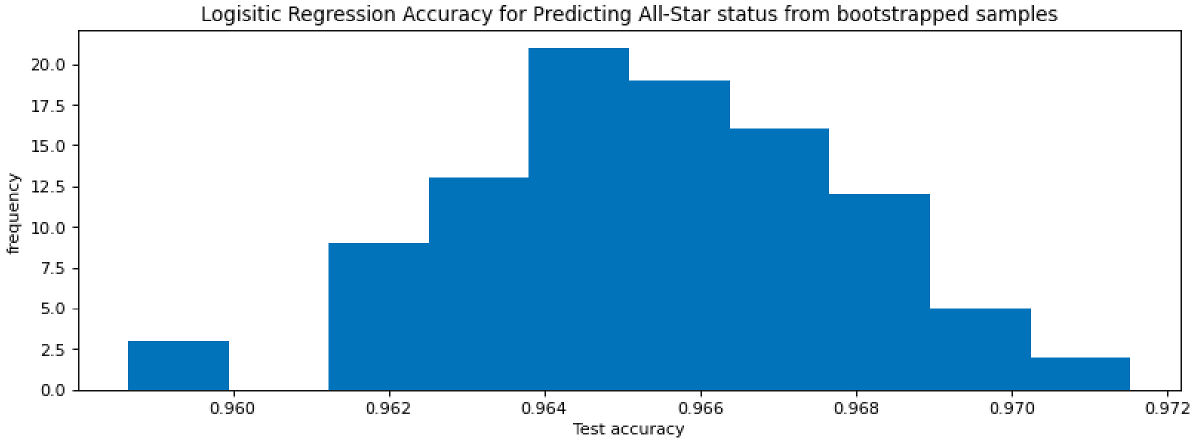
In order to estimate any uncertainty in our model, we decided to use the bootstrapping technique to predict our test and training accuracy on 100 samples of 80% of our data. We ended up getting a 95% confidence interval of [0.96500, 0.96548] for our training accuracy and [0.96453, 0.96559] for our test accuracy.

Training Accuracy Visualization:



Data C102 Checkpoint 2

Testing Accuracy Visualization:



Jonathan Wang <jonathan_wang@berkeley.edu>

Data 102 Final Project Proposal

1 message

Google Forms <forms-receipts-noreply@google.com> Wed, Nov 10, 2021 at 12:44 PM
To: jonathan_wang@berkeley.edu

Thanks for filling out Data 102 Final Project Proposal

Here's what was received.

Data 102 Final Project Proposal

Your team's project proposal is due by 11/10 at 11:59pm PST. Only one team member needs to complete this form.

Your email (**jonathan_wang@berkeley.edu**) was recorded when you submitted this form.

Team member #1's name *

Jonathan Wang

Team member #1's (UC Berkeley) email *

jonathan_wang@berkeley.edu

Team member #2's name: *

Rodrigo Cochran

Team member #2's (UC Berkeley) email *

rodrigocochran@berkeley.edu

Team member #3's name: *

Noah Kaminer

Team member #3's (UC Berkeley) email *

nrkaminer@berkeley.edu

Team member #4's name: *

NA

Team member #4's (UC Berkeley) email *

NA@gmail.com

Choice of dataset *

○ Chronic Disease and Air Quality

○ Transportation, Mobility, and Infrastructure

○ Primary Election Endorsements and Financing

● Other: The Lahman Baseball Database

If you selected an external dataset, you must include all relevant files in your project submission. Do you have permission to share your dataset with course staff? *

○ I am using a suggested dataset

● Yes, my external dataset is public

○ Yes, my external dataset is private but I have explicit permission to share it with course staff

Please list your two research questions.

Refer to the project specification for more details

Research Question #1 *

What is the causal effect of hitting statistics on a fielder's salary?

Which technique are you using to answer research question 1?

○ Binary decision-making

○ Bayesian Hierarchical Modeling

● Causal inference

○ Comparing GLMs and nonparametric methods

Research Question #2 *

Can we predict if a pitcher is going to be an all star in a given year based on pitching statistics?

Which technique are you using to answer research question 2?

● Binary decision-making

○ Bayesian Hierarchical Modeling

○ Causal inference

○ Comparing GLMs and nonparametric methods