Prabhat Nuthakki
Josh Geng

Data 102 Final Project Report

Data Overview

For our project analysis, we have chosen to work with the provided NBA dataset. This NBA dataset consists of five different tables that provide information statistics about the NBA season over the course of 14+ years. The dataset gives us information about factors like player statistics (PTS, MIN, etc.), team statistics (Location, FG%), and game statistics (game-winner, points scored by team). In order to make use of the data, as information is widely spread apart in the files, we had to manipulate the data in order to join tables and bring useful information together. Additionally, as explained in a later section, we made use of an external data set[1], NBA Injuries, that was available on Kaggle, providing us with valuable information regarding team injuries over the last decade, which we incorporated into our analysis.

The data was generated from a census rather than a sample as it covered all NBA games starting from 2014, with no groups excluded. Each row details individual player game-specific statistics, which will be aggregated for broader trend analysis.

Every participant within the NBA, implicitly or explicitly, consents to the data collection of their in-game performance, which is a standard and public aspect of professional sports analytics. As for granularity, each row in our dataset corresponds to an individual player's stat line in a specific game, allowing us to analyze performance on a game-by-game, player-by-player basis. This high level of detail will enable us to interpret trends and patterns with significant precision but necessitates careful aggregation to understand broader trends.

Regarding data concerns, we face no selection bias due to the census nature of the dataset, but we remain vigilant about potential measurement errors, which are minimized but not entirely absent in sports statistics. Convenience sampling is not a concern, as the dataset comprehensively covers all NBA games during the specified period. Additionally, since the dataset is a comprehensive and public record of the performance of players, the dataset was not modified for differential privacy.

When working with the dataset, a column including a metric of player performance rating would have been extremely helpful in answering our research questions as they revolve around how well a player is performing, which cannot always be measured by the amount of points scored as players have different roles in teams. To combat this, we created our own Player Efficiency

[1] https://www.kaggle.com/datasets/ghopkins/nba-injuries-2010-2018

Rating, where we took stats such as points scored, offensive rebounds, etc to measure how impactful a player is in game. Another column that would be helpful would be a column that shows each player's position that they play so that we can do analysis based off of position, however the dataset only provides information on the positions of players who start in the game, thus greatly reducing the amount of data that we have to work with. As a result, we chose not to analyze by position in our project.

The main columns that we utilized that had missing data were PTS and MIN. For both PTS and MIN, many of the missing values can be attributed to the player not being able to play that day, either due to the coach's decision or the player being injured. By creating a binary variable for the column COMMENT, we can first remove all the rows where a player did not play, and then proceed with the dataset.

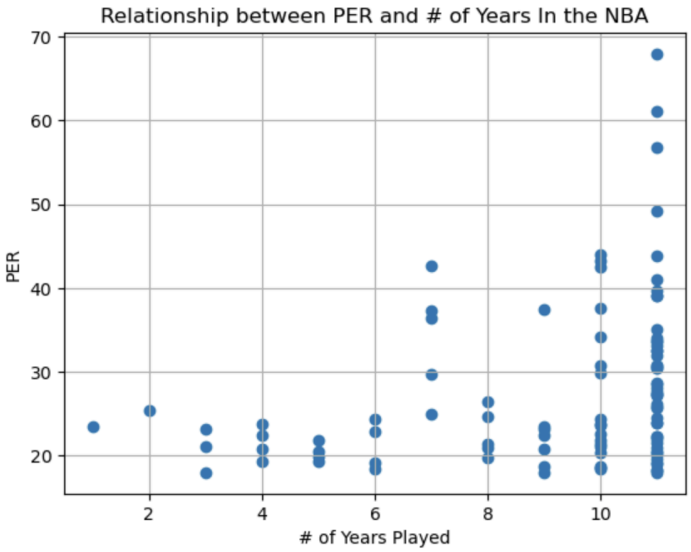
During data cleaning and preprocessing, we addressed inconsistencies and missing values by dropping the row, normalized data formats, and merging relevant information from separate tables. These steps were essential to prepare the dataset for analysis and to ensure the validity of our subsequent models and inferences. The decisions made during this process could potentially impact the outcome of our analyses as some rows of player data are omitted.

Order to Run Files:

1) **Data 102 Project Code Part 1**
2) **Data 102 Project Code Question #2**
3) **Data 102 Project Code Question #2 Part #2**

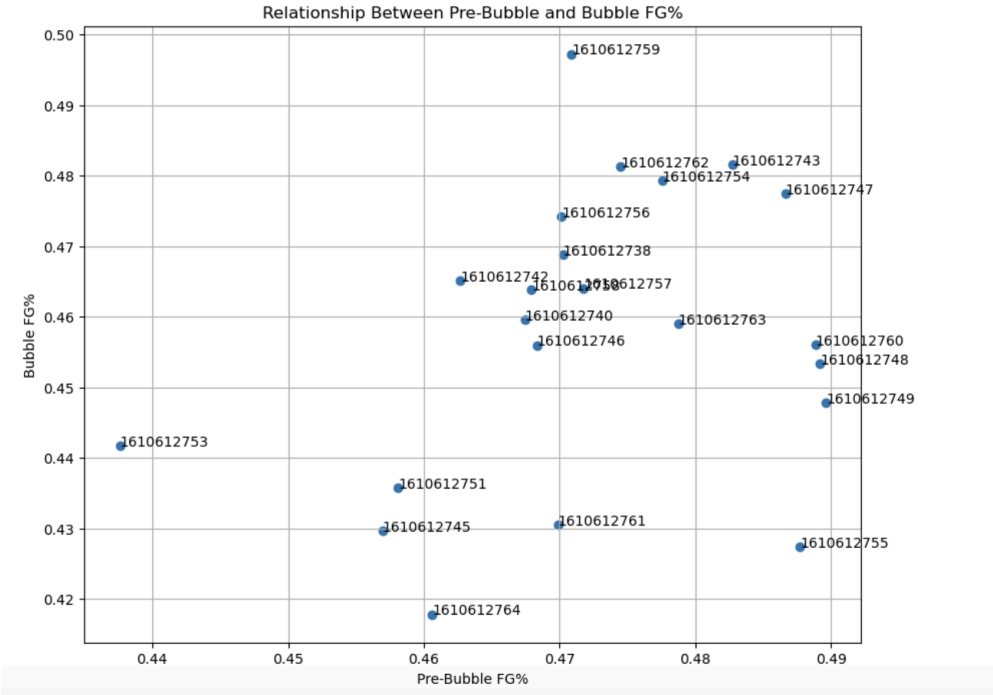
EDA

The EDA shown below portrays relationships between variables in the NBA dataset that we manipulated to make use of our analysis. We essentially try to show the relationship between player performance, compared to a metric Efficiency Rating, which we compiled for our use. We define Efficiency rating as the following formula (Points + Rebounds + Assists + Steals + Blocks - Turnovers) / (Total Minutes Played) across all game data given by the 'games_details.csv.' By taking the top 100 players in PER rating (who had at least 500 minutes played overall), we plotted it against the number of years a player has been playing in the league.



This visualization is relevant to our research question as it shows the dynamic between the experience of playing in the league and the player's performance in-game. The above graph suggests that the Player Efficiency Rating increases as a player spends more time in the league. we are able to see that most of the top-performing players in terms of PER are veterans, players who have played for 7 years or more. The highest PER players are ones who have played 11 years in the league, but some veterans rank at the bottom of the top 100 players. One potential follow-up would be how age affects PER, given that basketball is a physically intensive sport, and player skill would likely deteriorate with age after a certain point, thus resulting in a lower PER despite the number of years spent playing in the league.

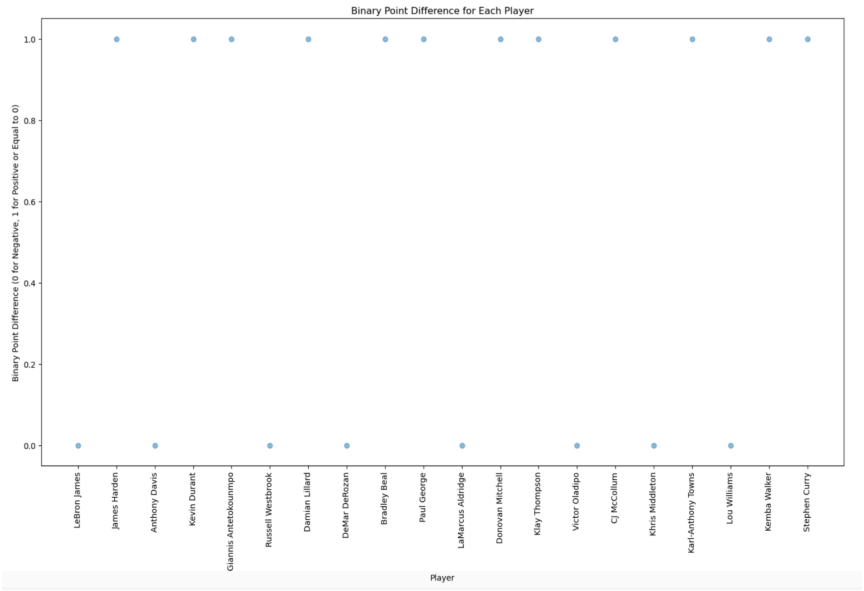
FG% for home games before the Bubble (Oct 22, 2019 - March 11, 2020) vs FG% for home games in the Bubble (July 22 - Oct 11, 2020)



This visualization is relevant to our research question because it shows how the scoring ability of a player is affected by playing in the Bubble where there is no live audience to give the home players a “buff” amongst many other factors such as unfamiliar environment, elongated break during the season, etc. According to the visualization, it appears that a majority of the teams fall below the Y=X line, which shows that these teams have a lower FG% playing in the bubble when compared to before playing in the bubble. A few teams fall significantly under the line, which shows that these teams are heavily impacted by playing in the bubble. Follow-ups to this visualization would be how playoff teams and the championship team were affected by playing in the bubble in terms of scoring and how top players were affected by playing in the bubble. Additionally, we could observe the difference in FG% for away games in and out of the bubble. The visualization suggests that playing in the bubble decreases a player’s ability to score points during home games.

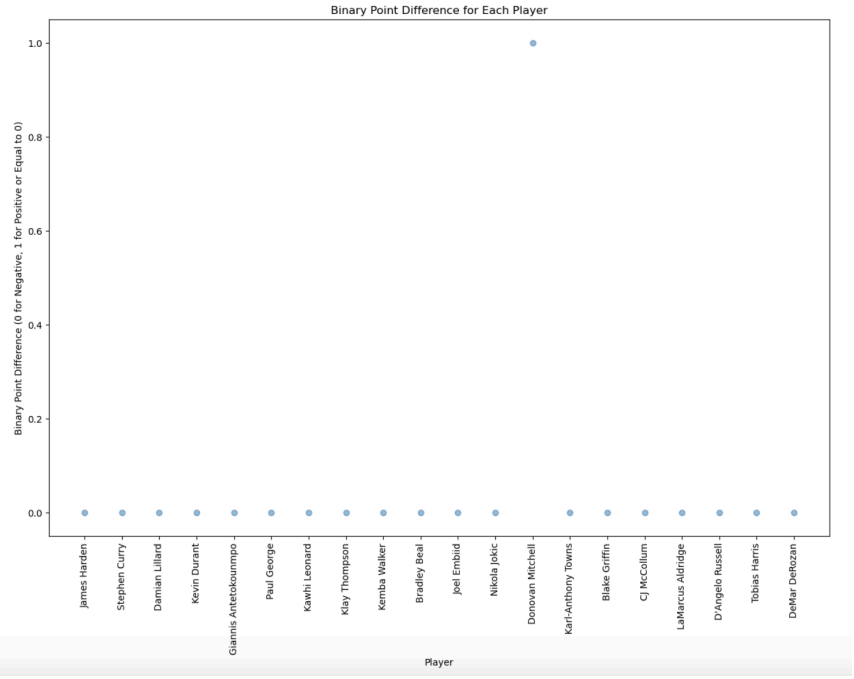
Points scored in the 2017-2018 Season compared to Points scored in the 2018-2019 Season (Playing Time, minutes)

SEASON / PLAYER_NAME	2017	2018	Point Difference	Binary Point Difference
LeBron James	3016.0	1560.0	-1456.0	0
James Harden	2762.0	3247.0	485.0	1
Anthony Davis	2459.0	1532.0	-927.0	0
Kevin Durant	2452.0	2486.0	34.0	1
Giannis Antetokounmpo	2235.0	2449.0	214.0	1
Russell Westbrook	2233.0	1789.0	-444.0	0
Damian Lillard	2132.0	2547.0	415.0	1
DeMar DeRozan	2113.0	1859.0	-254.0	0
Bradley Beal	2059.0	2167.0	108.0	1
Paul George	1941.0	2373.0	432.0	1
LaMarcus Aldridge	1941.0	1914.0	-27.0	0
Donovan Mitchell	1935.0	1982.0	47.0	1
Klay Thompson	1932.0	2203.0	271.0	1
Victor Oladipo	1923.0	734.0	-1189.0	0
CJ McCollum	1900.0	1923.0	23.0	1
Khris Middleton	1873.0	1711.0	-162.0	0
Karl-Anthony Towns	1866.0	1974.0	108.0	1
Lou Williams	1833.0	1658.0	-175.0	0
Kemba Walker	1832.0	2175.0	343.0	1
Stephen Curry	1811.0	2584.0	773.0	1



The Y-axis is binary, 1 if the player scored more in the 2018-2019 season, and 0 otherwise. The X-axis is the names of the 20 top players sorted by Points scored in the 2017-2018 season. This visualization is important because it shows how one additional year of experience playing in the league affects the scoring potential of top talents in the league. This helps give us an understanding of our second research question, showing that essentially more minutes played could have an impact on future scoring. Although this isn't in terms of minutes, playing an extra season does mean that a player has more playing time. Factors such as injuries, whether or not the team made it to the playoffs, or being on a different team altogether make for outliers and inconsistencies in the visualization.

Points scored in the 2018-2019 Season vs Points Scored in the Bubble Season



The Y-axis is binary, 1 if the player scored more in the bubble season, and 0 otherwise. The X-axis is the names of the 20 top players sorted by points scored in the 2018-2019 season. This visualization helps address the research question by providing a nuanced understanding of how individual players fared during the NBA bubble season compared to their previous season's performance, contributing valuable insights to the broader analysis of the impact of the bubble season on points scored.

Research Question

Part 1: How did playing in the NBA bubble (Jul 7, 2020 – Oct 11, 2020) affect the FG% of the teams that participated compared to the team's FG% the previous season?

In this analysis, the treatment is playing in the NBA bubble, which is represented by the date when the games were played (converted into a binary variable). We essentially had to filter out the table using the dates of the games, to only work with games that fell in between the period for the bubble. The outcome is the field goal percentage of the teams.

Confounders:

Team performance prior to playing in the bubble is also a confounding variable. If the team's performance is trending downwards near the end of the pre-bubble period, they could

potentially perform poorly whilst in the bubble. The unconfoundedness assumption holds here, as we can observe the trends of teams in regard to field goal percentage across the season. Injuries are a confounding factor, if a team has many injured players, the dynamics of the game change, and missing key players/defenders can greatly influence the field goal percentage of a team. We obtained the data for the confounding variable injuries through an external dataset, which provides NBA injury data over many seasons. The unconfoundedness assumption holds here as we obtained data on the injuries for both pre-bubble and bubble games.

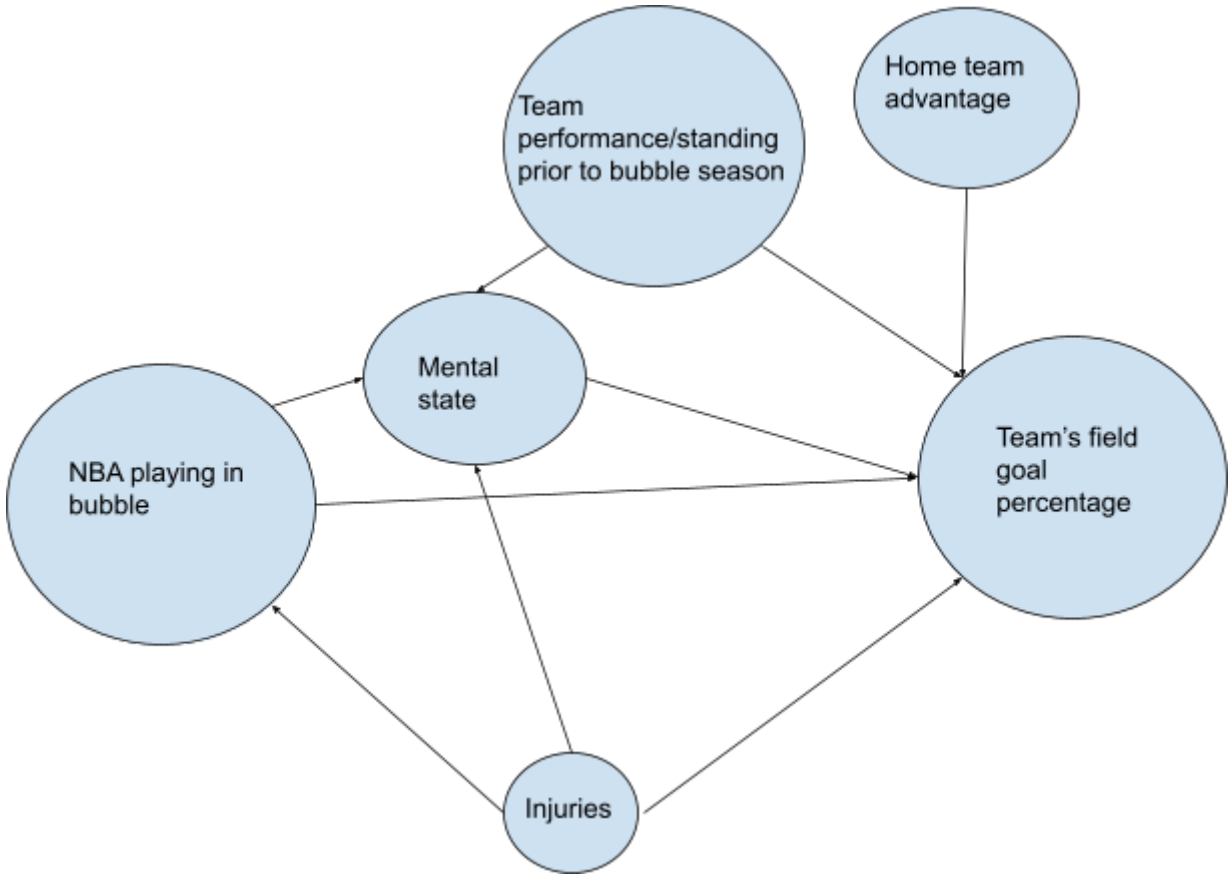
Home team advantage could also be a confounder as it is believed that teams generally perform better in home games compared to away games, but in the bubble, there were no fans and all games were played in the same location, thus there was effectively no home team advantage during the latter half of the season. The unconfoundedness assumption holds as we have data on who the home and away teams were in each game. The mental state of a player is another confounding factor as the games were being played during the brunt of the pandemic. Many things in life seemed uncertain, and many people were falling ill or had close friends or family affected by COVID, thus potentially hindering the performance of players. The unconfoundedness assumption does not hold here as we have no information about what a player might be going through personally. We have no viable metrics or data to test for this assumption.

To account for these confounders we will use sensitivity analysis and the inclusion of binary variables from obtained/manipulated data in order to incorporate feasible and significant confounders. For example, for the confounding variable, injuries, we were able to obtain external data in order to incorporate injury data for each of the teams. Similarly, we were able to incorporate a metric, home_game_advantage, to simulate the effect of home game advantages in relation to FG%. From the data we could find/extract in the NBA files, we were able to find out how many home games each of the teams that participated in the bubble played throughout the whole season. As we don't have metrics about the exact home games themselves, like the # of fans attended, we could only do so much with this. We found that the spread of home games is only from 33 - 37 for all teams, so although we chose 35 as a cutoff, it does not represent a significant value compared to one with a larger spread. We decided to still use this metric to see if it had any potential effect. This confounder, unlike injuries, from the data we have is a harder one to properly measure.

<u>Below is a list of potential colliders we found relating to our analysis.</u>

- Playing in the bubble and mental state both influencing field goal percentage
- Playing in the bubble and injuries both influencing field goal percentage
- Playing in the bubble and home team advantage both influencing field goal percentage
- Injuries and mental state both influencing field goal percentage
- Injuries and home team advantage both influencing field goal percentage
- Mental state and home team advantage both influencing field goal percentage

- Injuries and home team advantage both influencing field goal percentage
- Playing in the bubble and injuries both influencing mental state



OLS Regression Results

Dep. Variable:	FG_PCT_home	**R-squared:**	0.161
Model:	OLS	**Adj. R-squared:**	0.120
Method:	Least Squares	**F-statistic:**	3.926
Date:	Wed, 06 Dec 2023	**Prob (F-statistic):**	0.0275
Time:	03:17:13	**Log-Likelihood:**	117.46
No. Observations:	44	**AIC:**	-228.9
Df Residuals:	41	**BIC:**	-223.6
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.4692	0.011	44.502	0.000	0.448	0.490
C(period)[T.1]	-0.0124	0.009	-1.418	0.164	-0.030	0.005
injuries	0.0001	0.000	0.304	0.762	-0.001	0.001

This regression is examining whether there is a statistically significant difference in FG% between the pre-bubble and during-bubble periods after controlling for the number of injuries. The coefficient for c(period) will tell us the difference in FG% associated with the during-bubble period compared to the pre-bubble period. The coefficient for 'injuries' will tell us the association between the number of injuries and FG%. This model suggests that, based on the data

and the model specified, neither the change in periods (pre-bubble to during-bubble) nor the number of injuries has a statistically significant impact on FG%. Given the relatively low R-squared value, there may be other factors not included in the model that could explain variation. These other factors could be confounding variables that are not feasible to be measured but have a correlation. The estimated difference in FG% between the pre-bubble (period 0) and during-bubble (period 1). The coefficient is -0.0124, but it's not statistically significant (p-value is 0.164), suggesting that there is no strong evidence of a difference in FG% between the two periods. We can, not significantly, verify a small similarity between the two by looking at the difference in means of FG% for all teams before and during the bubble - Pre-Bubble: 0.4721738636363636, Bubble: 0.45758859090909093.

This regression is similar to the regression above, however, it accounts for FG% for the away teams instead of home. We cannot strongly suggest that a home game versus an away game had a meaningful impact on a team's FG% during this period.

OLS Regression Results

Dep. Variable:	FG_PCT_away	**R-squared:**	0.086
Model:	OLS	**Adj. R-squared:**	0.041
Method:	Least Squares	**F-statistic:**	1.920
Date:	Wed, 06 Dec 2023	**Prob (F-statistic):**	0.160
Time:	04:01:52	**Log-Likelihood:**	107.67
No. Observations:	44	**AIC:**	-209.3
Df Residuals:	41	**BIC:**	-204.0
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.4635	0.013	35.191	0.000	0.437	0.490
C(period)[T.1]	-0.0001	0.011	-0.013	0.989	-0.022	0.022
injuries	-0.0007	0.001	-1.180	0.245	-0.002	0.000

```
==============================================================================
Dep. Variable:     Bubble_FG_PCT_home   R-squared:                       0.240
Model:                            OLS   Adj. R-squared:                  0.061
Method:                 Least Squares   F-statistic:                     1.343
Date:                Tue, 05 Dec 2023   Prob (F-statistic):              0.295
Time:                        20:19:35   Log-Likelihood:                 57.664
No. Observations:                  22   AIC:                            -105.3
Df Residuals:                      17   BIC:                            -99.87
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                             coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept                  0.2432      0.169      1.436      0.169      -0.114       0.600
Pre_Bubble_FG_PCT_home     0.4981      0.363      1.371      0.188      -0.268       1.265
Pre_bubble_injuries       -0.0010      0.001     -1.574      0.134      -0.002       0.000
Bubble_injuries           -0.0004      0.001     -0.239      0.814      -0.003       0.003
home_field_advantage       0.0101      0.010      1.014      0.325      -0.011       0.031
------------------------------------------------------------------------------
```

This regression shows the relationship between a team's field goal percentage during the periods of pre-bubble to the bubble times using other factors like Pre bubble FG%, Pre bubble injuries, and home_field_advantage. For the Pre_bubble_FG_PCT_home, The coefficient is 0.4981, and the p-value is 0.188, which is above the 0.05 threshold for statistical significance. This suggests that there is a positive relationship between pre-bubble FG% and during-bubble FG%, but it is not statistically significant at the 5% level. For each one percentage point increase in pre-bubble FG%, the bubble FG% is predicted to increase by 0.4981 percentage points, if all other variables remain constant. For the relationship between the coefficients for Pre_bubble_FG%_home and home_field_advantage, the regression suggests that the pre-bubble FG% has a larger (but not statistically significant) impact on the FG% during the bubble than the home-field advantage does. The home-field advantage seems to have a very small and non-significant positive effect on the FG% during the bubble.

Discussion:

Basketball is an extremely complex sport that has been evolving for years, and the NBA is where the highest level of competition takes place. That being said, many factors affect a player's performance, and many are considered intangibles from a quantitative standpoint. This just so happens to be the greatest limitation that our method faces. There are simply too many factors that can hinder players, and things such as teamwork, interpersonal relationships within the team, and mental factors are all potentially confounding variables that we have no real way to incorporate into our analysis. Thus a big limitation of our method is the intangible confounding variables that we cannot quantify. During COVID where people around the world are all adapting differently to the new challenges that arose, the players are also affected in differing magnitudes. This is just one example of how ecological fallacy is also a limitation of our method, where things that we assume about the group may not be true about individuals. Lastly, we chose to limit our analysis to the top players, which could result in selection bias, which is yet another limitation of our methods, however, by only looking at top players, some of the intangibles such as choking under pressure are partially mitigated, as the top players are somewhat all battle-tested.

Additional data such as psychological and mental health assessments throughout the 2019 season would be useful to answering the causal question as we would be able to see how psychological and mental health changed throughout the season. This information would then help us understand the impact that playing in the bubble had on players' mental state, which allows us to then use that as a factor in our regression. Quantifying additional variables such as player condition or how they feel about the environment that they are in for the 2019 season would also be useful for the same reasons.

Due to our findings from our regression results, we cannot be confident in the relationship between playing in the bubble and player performance. None of the factors returned with a

statistically significant p-value, which leaves us rejecting the causal relationship. Additionally, there are too many limitations such as unobserved variables that we cannot be confident in our claim.

Part 2: BDM and Hypothesis Testing, How do in-game performance metrics - scoring, turnovers, efficiency - relate to a player's playing time, varying with location?

Our overall hypothesis is that players who are on the court for more minutes result in higher performance metrics. The reason why we are performing a multiple hypothesis test is to check multiple metrics of performance such as scoring, PER, and field goal percentage. Additionally, we check to see if factors such as playing in different cities or having a higher number of turnovers in a game will affect the player's performance.

The six hypotheses are listed below:

Relationship between MIN and PTS

Null: The amount of points scored is not influenced by the number of minutes played in the game.

Alternative: The amount of points scored differs by the number of minutes played in the game.

Relationship between MIN and CITY on PTS

Null: The amount of points scored is not influenced by the number of minutes played in a specific city they are playing in.

Relationship of MIN and PTS on CITY

Null: The amount of minutes played and points scored is not influenced by the city they are playing in.

Relationship between MIN and PER

Null: There is no significant relationship between playing time (MIN) and player efficiency rating (PER).

Relationship of MIN and TO on PER

Null: There is no significant relationship between playing time (MIN) and turnovers (TO) on player efficiency rating (PER).

Relationship between MIN and FG_PCT

Null: There is no significant relationship between playing time (MIN) and field goal percentage (FG_PCT).a

Testing multiple hypotheses instead of just one allows for a better understanding of the nature of player performance in basketball. By exploring more than one, we can understand how different variables interact with each other, and how they might affect others. We can get a stronger sense of how closely they are associated with one another. In the NBA, and basketball generally, there are many components and metrics used to quantify a 'good player', and multiple hypotheses allow us to test on a broader range of metrics.

In order to test our hypotheses, we will be using correlation/association. Since the variables in this analysis are continuous, correlation tests are better suited to measure the direction of a linear relationship between two continuous variables. Correlation can be applicable in a case like this where our goal is to analyze multiple pairwise relationships. This is suitable for our scenario because many interrelated factors affect the outcome - player performance. Since all this analysis is on pre-existing data, association/correlation can be used to identify relationships. In 5/6 hypotheses, we are testing on some metric of player performance either in the form of PER or PTS as the outcome.

The two ways that we will correct for our multiple hypothesis testing are Bonferroni Correction and False Discovery Rate correction. The Bonferroni correction method will help us to control the family-wise error rate to ensure that our probability of making type I errors across our tests is under our chosen significance level. The FDR correction method will control the expected amount of rejected hypotheses being marked as true.

<u>Relationship between MIN and PTS</u>

```
Discoveries with FWER control (Bonferroni): 2
Discoveries with FDR control (Benjamini-Hochberg): 2
                            OLS Regression Results
==============================================================================
Dep. Variable:                    PTS   R-squared:                       0.357
Model:                            OLS   Adj. R-squared:                  0.357
Method:                 Least Squares   F-statistic:                 1.316e+04
Date:                Mon, 11 Dec 2023   Prob (F-statistic):               0.00
Time:                        19:34:35   Log-Likelihood:                -71040.
No. Observations:               23661   AIC:                         1.421e+05
Df Residuals:                   23659   BIC:                         1.421e+05
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.4355      0.071      6.106      0.000       0.296       0.575
MIN            0.3945      0.003    114.734      0.000       0.388       0.401
==============================================================================
```

The 'MIN' coefficient is 0.3945, meaning that for each additional minute played, the total points scored (PTS) are expected to increase by 0.3945 points on average. This relationship is also highly statistically significant, suggesting a strong positive relationship between playing

time and scoring. Both the FWER and FDR yielded 2 each, meaning that when adjusting for the risk of Type 1 errors, there were still 2 statistically significant findings. Since they yielded the same discoveries, there is no leading to a more conservative result - no multiples here. When calculating the power for this test we received a 1 or (100%). A power of 1 suggests that the sample size is more than adequate to detect the observed relationship between MIN and PTS as statistically significant. Although a 1 in the real world is unlikely to achieve, we believe that due to rounding errors in the dataset, it should be a number close to 1. Given the regression results, where the p-value for MIN is indicating a significant result, and assuming the power calculation is very close to one, we would reject the null hypothesis in favor of the alternative. This would suggest a high level of confidence in the finding that the number of minutes played has a statistically significant influence on the number of points scored.

Relationship between MIN and CITY on PTS

```
Number of discoveries controlling FWER (Bonferroni): 2
Number of discoveries controlling FDR (Benjamini-Hochberg): 2
                            OLS Regression Results
==============================================================================
Dep. Variable:                    PTS   R-squared:                       0.360
Model:                            OLS   Adj. R-squared:                  0.359
Method:                 Least Squares   F-statistic:                     428.7
Date:                Mon, 11 Dec 2023   Prob (F-statistic):               0.00
Time:                        19:51:25   Log-Likelihood:                -70994.
No. Observations:               23661   AIC:                         1.421e+05
Df Residuals:                   23629   BIC:                         1.423e+05
Df Model:                          31
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.2356      0.188      1.250      0.211      -0.134       0.605
MIN            0.3946      0.003    114.434      0.000       0.388       0.401
Boston         0.0836      0.241      0.346      0.729      -0.390       0.557
Brooklyn      -0.1332      0.417     -0.320      0.749      -0.950       0.684
Charlotte     -0.5406      0.242     -2.230      0.026      -1.016      -0.065
Chicago       -0.0139      0.245     -0.057      0.955      -0.493       0.465
Cleveland     -0.1845      0.246     -0.749      0.454      -0.667       0.298
Dallas         0.2101      0.242      0.869      0.385      -0.264       0.684
Denver         1.0336      0.250      4.142      0.000       0.545       1.523
Detroit        0.1315      0.241      0.545      0.586      -0.342       0.605
Golden State   0.4596      0.253      1.814      0.070      -0.037       0.956
Houston        0.4392      0.250      1.757      0.079      -0.051       0.929
Indiana        0.3536      0.242      1.459      0.145      -0.121       0.828
Los Angeles    0.2483      0.214      1.163      0.245      -0.170       0.667
Memphis       -0.0572      0.250     -0.229      0.819      -0.548       0.433
Miami         -0.1388      0.248     -0.560      0.575      -0.624       0.347
Milwaukee     -0.2556      0.245     -1.043      0.297      -0.736       0.225
```

The coefficient for 'MIN' is 0.3946, indicating that for every additional minute played, the points scored increase by an average of 0.3946 points. This effect is statistically significant. Each team city has a coefficient representing its effect on the points scored, compared to the omitted reference city. The coefficient for Boston is 0.0836, but with a p-value of 0.729, it's not statistically significant, meaning that playing in Boston does not have a statistically significant effect on points scored compared to the reference city. The coefficient is -0.5406 with a p-value of 0.026, which is statistically significant. This suggests that playing for Charlotte is associated with a decrease in points scored by 0.5406 points compared to the reference city. Since there is essentially no control variable or control city to which metrics can be compared, metrics are analyzed compared to the default reference point (the first city in our dataset).

Relationship of MIN and PTS on CITY

```
Number of discoveries controlling FWER (Bonferroni): 7
Number of discoveries controlling FDR (Benjamini-Hochberg): 18
                            OLS Regression Results
==============================================================================
Dep. Variable:                    PTS   R-squared:                       0.363
Model:                            OLS   Adj. R-squared:                  0.361
Method:                 Least Squares   F-statistic:                     220.0
Date:                Mon, 11 Dec 2023   Prob (F-statistic):               0.00
Time:                        19:58:57   Log-Likelihood:                -70947.
No. Observations:               23661   AIC:                         1.420e+05
Df Residuals:                   23599   BIC:                         1.425e+05
Df Model:                          61
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975
------------------------------------------------------------------------------
const         -0.6364      0.399     -1.595      0.111      -1.418       0.14
MIN            0.4392      0.018     23.937      0.000       0.403       0.47
Boston         0.4993      0.534      0.935      0.350      -0.547       1.54
Brooklyn      -0.8814      1.081     -0.816      0.415      -3.000       1.23
Charlotte      1.8639      0.527      3.537      0.000       0.831       2.89
Chicago        0.7580      0.556      1.363      0.173      -0.332       1.84
Cleveland      1.1964      0.574      2.084      0.037       0.071       2.32
Dallas         1.7800      0.554      3.216      0.001       0.695       2.86
Denver         2.1829      0.579      3.769      0.000       1.048       3.31
Detroit        1.4505      0.539      2.693      0.007       0.395       2.50
Golden State   0.7338      0.535      1.372      0.170      -0.315       1.78
Houston        2.3314      0.575      4.055      0.000       1.204       3.45
Indiana        1.1114      0.537      2.068      0.039       0.058       2.16
Los Angeles    0.0479      0.484      0.099      0.921      -0.901       0.99
Memphis        0.8346      0.587      1.421      0.155      -0.316       1.98
Miami          1.1893      0.560      2.124      0.034       0.092       2.28
MIN_x_Boston          -0.0188      0.026     -0.734      0.463      -0.069       0.031
MIN_x_Brooklyn         0.0395      0.052      0.764      0.445      -0.062       0.141
MIN_x_Charlotte       -0.1244      0.024     -5.154      0.000      -0.172      -0.077
MIN_x_Chicago         -0.0394      0.026     -1.527      0.127      -0.090       0.011
MIN_x_Cleveland       -0.0727      0.028     -2.631      0.009      -0.127      -0.019
MIN_x_Dallas          -0.0836      0.027     -3.139      0.002      -0.136      -0.031
MIN_x_Denver          -0.0589      0.027     -2.198      0.028      -0.111      -0.006
MIN_x_Detroit         -0.0688      0.025     -2.722      0.006      -0.118      -0.019
MIN_x_Golden State    -0.0136      0.024     -0.560      0.575      -0.061       0.034
MIN_x_Houston         -0.0995      0.027     -3.655      0.000      -0.153      -0.046
MIN_x_Indiana         -0.0383      0.026     -1.498      0.134      -0.088       0.012
MIN_x_Los Angeles      0.0139      0.023      0.610      0.542      -0.031       0.058
MIN_x_Memphis         -0.0457      0.027     -1.670      0.095      -0.099       0.008
MIN_x_Miami           -0.0695      0.027     -2.619      0.009      -0.121      -0.017
MIN_x_Milwaukee       -0.0832      0.025     -3.282      0.001      -0.133      -0.034
MIN_x_Minnesota       -0.0443      0.027     -1.664      0.096      -0.096       0.008
MIN_x_New Jersey      -0.0226      0.027     -0.831      0.406      -0.076       0.031
MIN_x_New Orleans     -0.0364      0.027     -1.341      0.180      -0.090       0.017
MIN_x_New York        -0.0346      0.026     -1.314      0.189      -0.086       0.017
MIN_x_Oklahoma City   -0.0077      0.033     -0.237      0.813      -0.072       0.056
MIN_x_Orlando         -0.0243      0.027     -0.898      0.369      -0.077       0.029
MIN_x_Philadelphia    -0.0164      0.026     -0.621      0.535      -0.068       0.035
MIN_x_Phoenix         -0.0841      0.027     -3.103      0.002      -0.137      -0.031
MIN_x_Portland        -0.0490      0.026     -1.908      0.056      -0.099       0.001
MIN_x_Sacramento      -0.0265      0.025     -1.040      0.298      -0.076       0.023
MIN_x_San Antonio     -0.0553      0.032     -1.753      0.080      -0.117       0.007
MIN_x_Seattle         -0.0472      0.034     -1.405      0.160      -0.113       0.019
MIN_x_Toronto         -0.0539      0.028     -1.953      0.051      -0.108       0.000
MIN_x_Utah            -0.0355      0.027     -1.323      0.186      -0.088       0.017
MIN_x_Washington      -0.0460      0.027     -1.698      0.089      -0.099       0.007
```

The coefficient for 'MIN' (0.4392) is the effect of minutes played on points scored for the reference city (the city that was not included as a dummy variable, the first choice in the set). It indicates that, on average, for each additional minute played, there is an increase of 0.4392 points in the reference city. This effect is statistically significant. The interaction terms (with the 'MIN' prefix) are coefficients that represent how the effect of each additional minute played on points scored varies by city compared to the reference city. For example, The coefficient of Charlotte (-0.1244) suggests that the positive effect of each additional minute played on points scored is 0.1244 points less in Charlotte compared to the reference city. This effect is significant but for a city like Milwaukee, we can see that it is not, because of the p-value. The negative sign on these interaction terms indicates that the additional points per minute gained in these cities are less than the points gained per minute in the reference city. The Number of discoveries controlling FWER (Bonferroni) is 7, which means after applying the conservative Bonferroni correction to control the family-wise error rate, there are 7 coefficients in the model that are statistically significant. The Number of discoveries controlling FDR is 18, which means when controlling the false discovery rate, which is less conservative than Bonferroni, there are 18 coefficients considered statistically significant. The differences between the FWER and FDR results indicate that while there are several significant predictors of points scored when we account for the risk of making any false discoveries across all tests (FWER), fewer predictors are considered significant. This could be relevant for strategic decisions, such as if a team's strategy should adjust when playing away games versus home games.

Relationship between MIN and PER

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    PER   R-squared:                       0.049
Model:                            OLS   Adj. R-squared:                  0.040
Method:                 Least Squares   F-statistic:                     5.099
Date:                Mon, 11 Dec 2023   Prob (F-statistic):             0.0262
Time:                        16:23:52   Log-Likelihood:                -365.01
No. Observations:                 100   AIC:                             734.0
Df Residuals:                      98   BIC:                             739.2
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         34.8849      3.350     10.413      0.000      28.237      41.533
MIN           -0.0081      0.004     -2.258      0.026      -0.015      -0.001
==============================================================================
Omnibus:                       43.791   Durbin-Watson:                   0.115
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               98.898
Skew:                           1.703   Prob(JB):                     3.35e-22
Kurtosis:                       6.483   Cond. No.                     3.32e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.32e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
Discoveries with FWER control (Bonferroni): 1
Discoveries with FDR control (Benjamini-Hochberg): 2
```

The MIN coefficient (-0.0081) represents the change in PER for each additional minute played. It's negative, suggesting that as players play more minutes, their efficiency (as measured by PER) slightly decreases. The p-value (0.026) indicates this result is statistically significant at the 5% level, but it is close to the cutoff. The R-squared value (0.049) indicates that about 4.9% of the variance in PER can be explained by the number of minutes played. This is a relatively low value, suggesting that minutes played only explain a small portion of the variation in PER. The result suggests that there is a small but statistically significant negative relationship between minutes played and player efficiency. This could imply that as players spend more time on the court, their overall efficiency might decrease, perhaps due to fatigue or other factors. The FWER control result being lower (only 1 discovery) than the FDR result (2 discoveries) indicates that when we're very conservative and want to minimize the chance of any false positives, we can conclude that 'MIN' in this case, is significantly related to PER.

Relationship of MIN and TO on PER

```
Discoveries with FWER control (Bonferroni): 3
Discoveries with FDR control (Benjamini-Hochberg): 3
                            OLS Regression Results
==============================================================================
Dep. Variable:                    PER   R-squared:                       0.772
Model:                            OLS   Adj. R-squared:                  0.768
Method:                 Least Squares   F-statistic:                     164.5
Date:                Mon, 11 Dec 2023   Prob (F-statistic):           6.87e-32
Time:                        16:33:17   Log-Likelihood:                -293.57
No. Observations:                 100   AIC:                             593.1
Df Residuals:                      97   BIC:                             600.9
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         26.8022      1.711     15.661      0.000      23.406      30.199
MIN           -0.0209      0.002    -10.943      0.000      -0.025      -0.017
TO             0.0098      0.001     17.545      0.000       0.009       0.011
==============================================================================
```

The table indicates that there are 3 discoveries when controlling for FWER with the Bonferroni method and 3 discoveries when controlling for FDR with the BH method. This means that after correcting for multiple tests, both the number of minutes played and turnovers remain statistically significant in predicting PER. However, this does

not provide a valuable conclusion because based on intuition, an increase in TO would decrease PER. However, this positive relationship might be due to a more complex dynamic. For instance, players who handle the ball more often (and therefore have more opportunities for turnovers) might also be more involved in plays that lead to scoring, thus having higher PERs. The negative coefficient for minutes played suggests that there might be diminishing returns in terms of efficiency as players play more minutes. This could be due to factors like fatigue.

OLS Regression Results

Dep. Variable:	FG_PCT	R-squared:	0.038
Model:	OLS	Adj. R-squared:	0.038
Method:	Least Squares	F-statistic:	941.0
Date:	Mon, 11 Dec 2023	Prob (F-statistic):	1.12e-202
Time:	17:07:57	Log-Likelihood:	-1309.4
No. Observations:	23661	AIC:	2623.
Df Residuals:	23659	BIC:	2639.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
tput; double click to hide						
const	0.3031	0.004	80.956	0.000	0.296	0.310
MIN	0.0055	0.000	30.675	0.000	0.005	0.006

Omnibus:	878.122	Durbin-Watson:	1.980
Prob(Omnibus):	0.000	Jarque-Bera (JB):	978.892
Skew:	0.497	Prob(JB):	2.73e-213
Kurtosis:	3.058	Cond. No.	46.8

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
Discoveries with FWER control (Bonferroni): 2
Discoveries with FDR control (Benjamini-Hochberg): 2

Relationship between MIN and FG_PCT

The coefficient for MIN is 0.0055, which means that for each additional minute played, the FG_PCT is expected to increase by an average of 0.55 percentage points. This is also statistically significant (p-value = 0.000), suggesting a positive relationship between playing time and shooting efficiency. The 2 discoveries for both FWER and FDR mean that the positive relationship between MIN and FG_PCT holds true after correcting for multiple tests.

Discussion

Although we were able to observe a positive correlation between the player's in-game performance and the number of minutes played after testing our six hypotheses, there could have been more factors that limited our analysis. For example, when there's a big lead during the game, the winning team tends to have the starting players get some rest and instead put benched players as so-called garbage time which could make the minutes played less effective in testing. Regardless, our hypothesis testing results were overall significant after applying Bonferroni and BH methods. It can be concluded that the more minutes played increased both points scored and the field goal percentage, however, the player's efficiency slightly decreased from the individual tests. If more data were provided, we could test for the relationship between player efficiency and the position they play, as well as the team's performance with and without a star player. Another test would be if the amount of points allowed by the team's defense affects the amount of points that the team's offense scores.

Examples of Potential Decisions:

MIN and PTS: Recognizing the direct relationship between minutes played and points scored,

coaches can make decisions about player rotation and playing time. Players who demonstrate the ability to maintain or increase scoring efficiency with more minutes might be given priority in critical game periods.

MIN and PER: Insights into how playing time affects player efficiency can inform decisions on player endurance training and identifying when players may be reaching the point of diminishing returns in their performance during a game.

Game Strategy: Understanding the nuances of how various factors interact with performance can refine game strategies, making them more adaptable to the context of each game.

Conclusion

We found that the NBA is an environment that contains many unobservable variables, which ultimately led to us not being confident in our claim that players performed worse while playing in the NBA bubble. We also found that there is a positive correlation between minutes played and points scored, however, their player efficiency rating tends to drop. However, even this analysis can be subject to the different decisions that people on the teams make, such as if a coach decides to tell their star player to step off the court and rest. Our results are not very generalizable, once again due to the amount of variables that basketball as a sport includes.

We merged the player injury data set with the provided NBA dataset, which allowed us to do analysis on how player injuries affect the performance of a team. This was a relatively smooth process as both datasets included the data and were very comprehensive.

There were no real limitations in the data as it included everything that it should have as an NBA stat sheet. Unfortunately, we were unable to find data on factors such as player mental health throughout the 2019 season from external sources that would have helped us in our analysis of the NBA bubble. Future studies such as the optimal amount of time for a player to play based on their player efficiency rating could be built off of our work, as we showed that points scored increase with minutes played, but player efficiency rating drops as a result of fatigue. This could help teams find a baseline for bench rotations, thus maximizing the scoring potential of the team.