Data Overview

We are using the NBA games dataset, 2004-2022 (Kaggle). This data is provided by the National Basketball Association, which provides detailed information about the teams, players, and games played at the national level for the past nearly two decades. It is neither a sample nor a census and also includes statistics calculated from the raw data collected. In addition to this dataset, we are also using the 20 Years of NBA Draft dataset (Kaggle, https://www.kaggle.com/datasets/benwieland/nba-draft-data/) to help us answer our second research question regarding predicting NBA drafts. This dataset includes a list of all players drafted into the NBA, data on which player was selected in the NBA draft, as well as some base statistics and background information of each player.

NBA players are aware of data collection and its use for broadcasting and prediction, as this information is part of their professional environment. They consent to this as part of their remuneration for playing. The data, including game recordings and statistics, is publicly available on the NBA website, ensuring transparency and accessibility. The specific details of each dataset are outlined in the following sections.

Players.csv: This csv has a low granularity as it only contains basic information about each player, such as the team id, player id, and season played. Each row represents a different player. This does not impact the interpretation of findings because player id, team id, and the season played do not affect our research question.

Ranking.csv: This csv has a high granularity as it includes everything from 2003 to 2022 in terms of standing dates, and includes detailed metrics such as win percentage. Each row represents a team in a certain year. So although there might be duplicates of teams, the year participated in the NBA might be different, hence a new row.

Games_details.csv: This csv has a high granularity, as it includes player statistics for each game, scores, and other game-specific details such as the team they belong to and where the team is from. Each row represents a new player within the team.

Draft-data-20-years.csv: This csv has high granularity, because it includes records for all drafted players in the past 20 years. However, this dataset, compared to the information on https://www.nba.com/stats/players/bio, does not include features such as height, weight, or age, which might be instrumental in the decision process of draft picking.

In our study, selection bias isn't a concern since we included comprehensive data from all relevant NBA teams and players within our timeframe. However, some measurement errors may exist due to the subjective nature of recording assists, points, rebounds, playtime, and fouls. Despite this, the data's reliability is high, sourced from the official NBA website. Also, convenience sampling was not a factor, as the entire population of relevant statistics was used, negating the need for sampling methods.

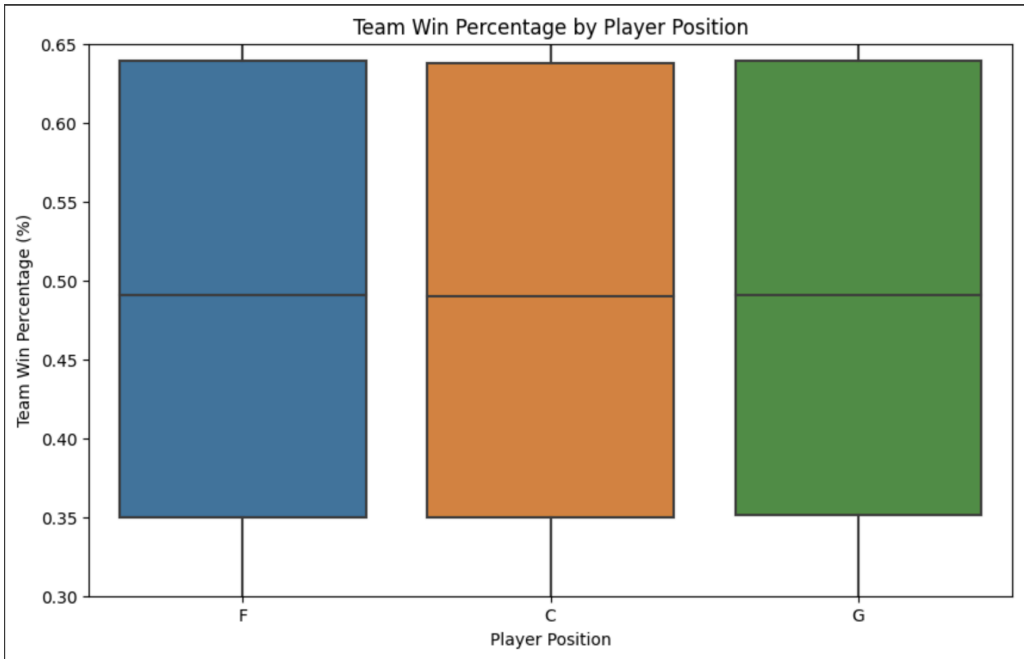
Our datasets weren't altered for differential privacy, as they involve public figures. Additional useful data would include each player's percentage contribution to the team's score per game, which would aid in assessing their impact on team wins. Including players' height, weight, and age could also enhance draft order predictions.

In "games_details.csv," missing data was handled by removing rows with NaN values, which represented unrecorded game data or inactive players. "Nickname" columns, often inapplicable, contributed to NaN values. For "draft-data-20-years.csv," NA indicated non-participation in NBA games, and zeroes represented lack of recorded stats. We assigned zeroes to missing values for consistency.

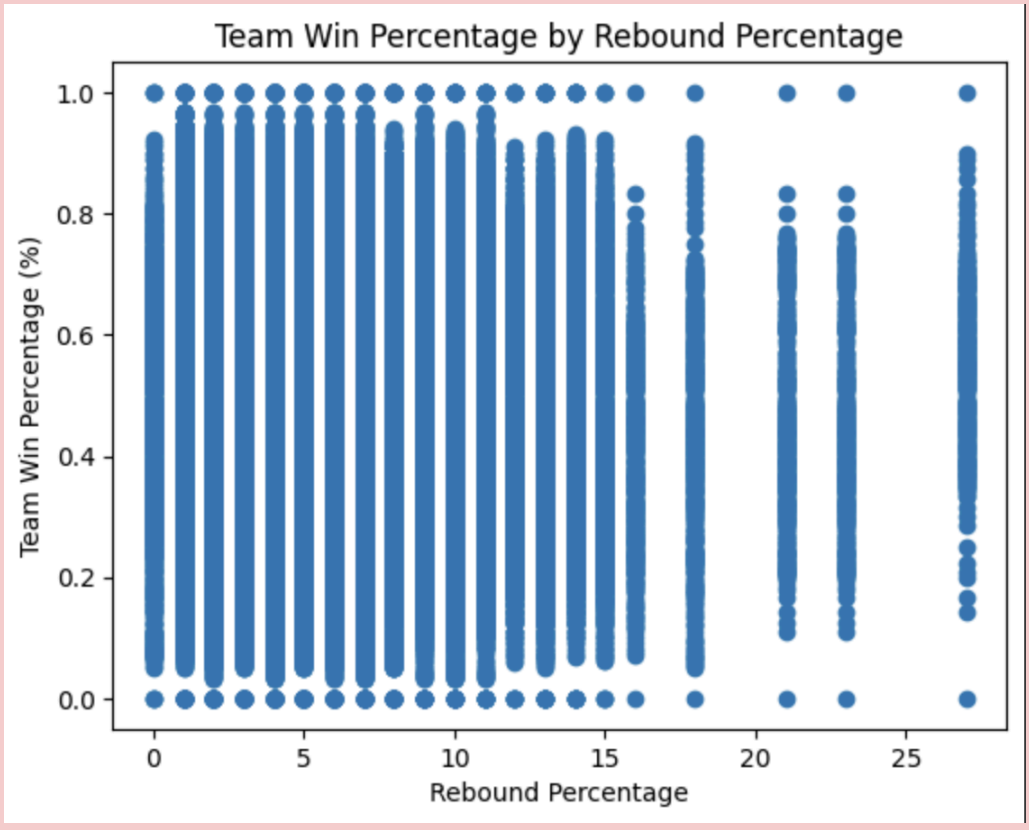
Unnecessary columns, like player names in "games_details.csv" and duplicate columns in "draft-data-20-years.csv," were removed to streamline the datasets for our models.

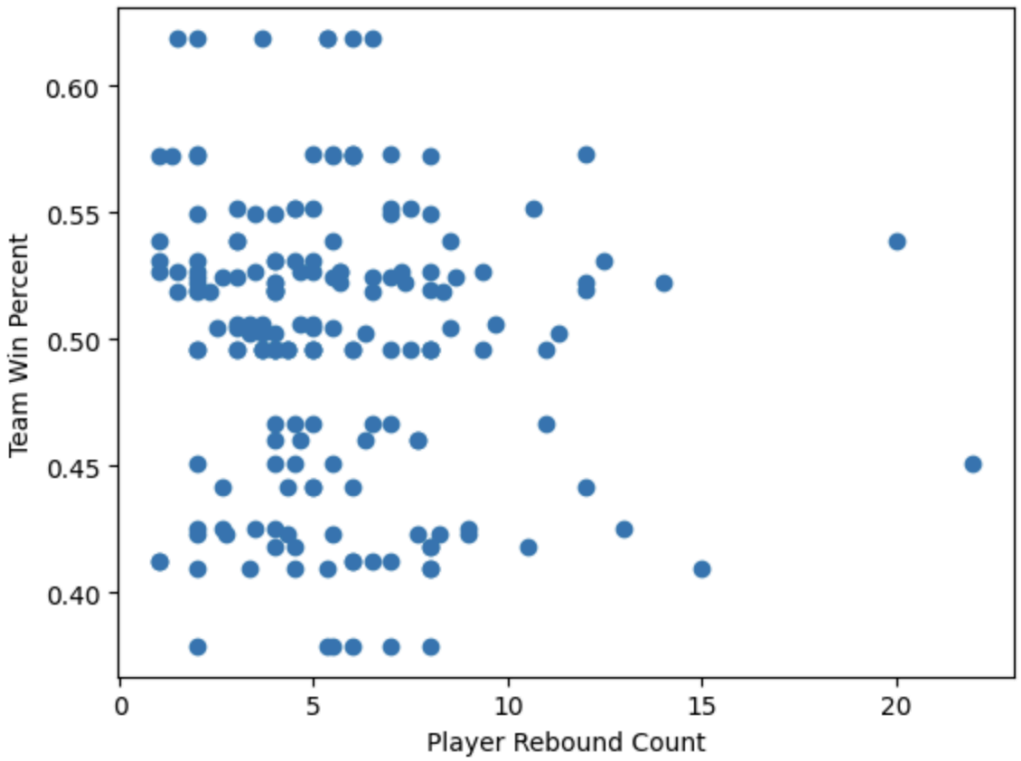
EDA

Research Question 1:



There seems to be an extremely small, almost imperceivable, difference between a player's position and their team's win percentages. Thus, the player's position is not a good way to predict their team's win percentage.

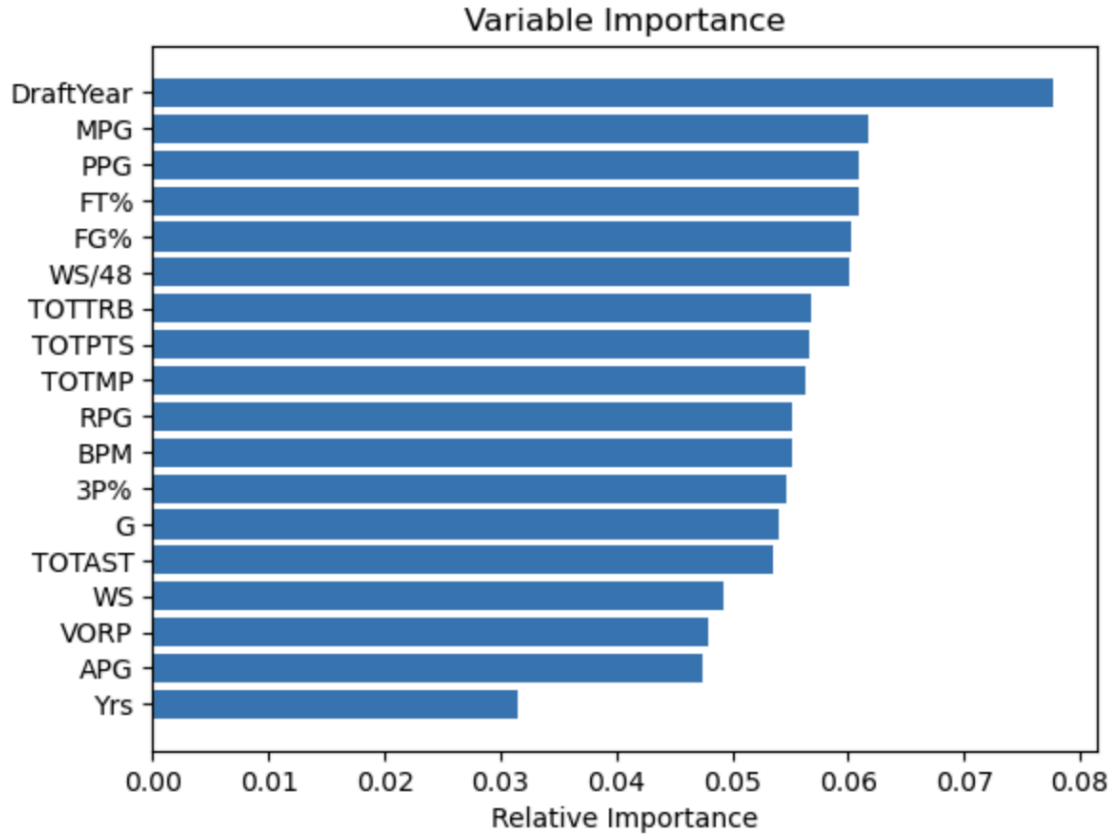




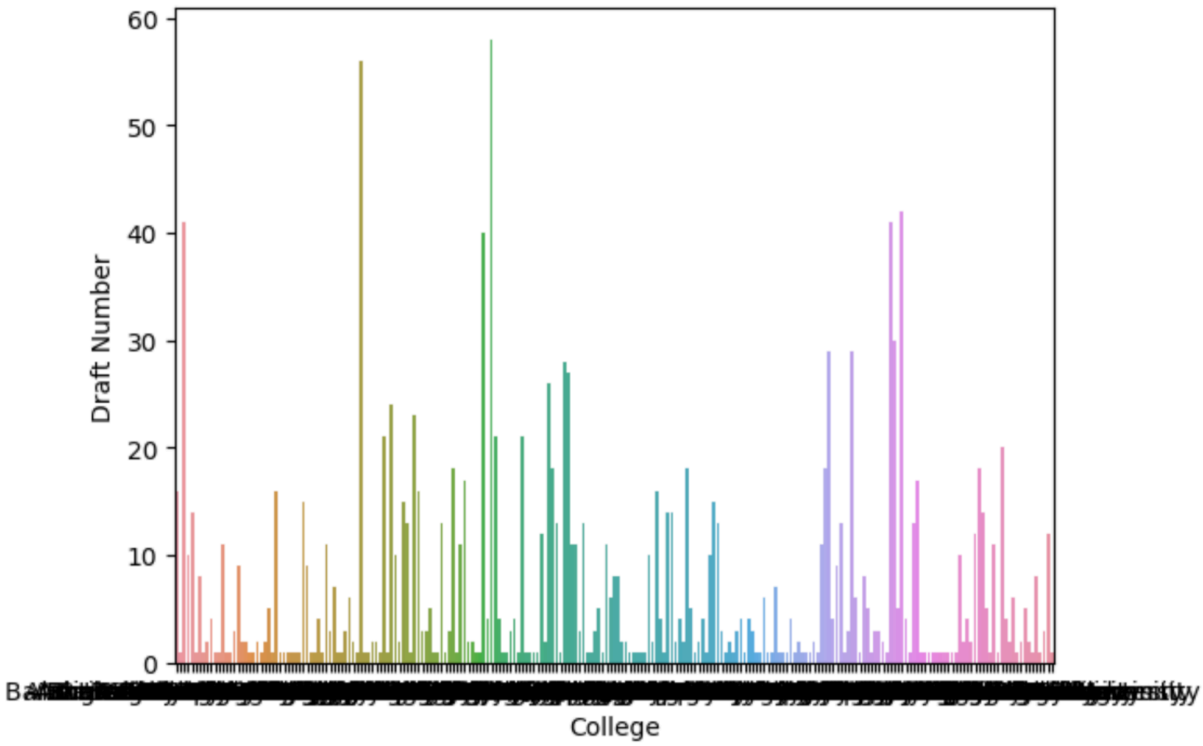
According to the plot, there is a low correlation between the rebound count and a team's win percentage. At a lower rebound count, the team's win percentage seems to be anywhere from .3 to .65 while there don't seem to be many players who get a high rebound count, and if they do there still seems to be no correlation with the team's win percentage. This suggests that using a player's past rebound count cannot be used to predict a team's win percentage.

Our findings suggest that player statistics, both categorical and quantitative, are not effective predictors of a team's win percentage. While these stats contribute to a team's score, they don't reliably indicate winning outcomes. Future research will explore other player statistics to identify variables with a stronger impact on win percentage.

Research Question 2:



To choose which features to use in the Random Forest Classifier, we must use feature engineering to sort the different features/columns from most to least relative importance. From the above graph, we can see that the most important features that can be used to predict rank are DraftYr, MPG, PPG, FT%, FG%, etc. This method is only able to rank the quantitative features. Some observations we can see is that DraftYear seems to have the most relative importance, by a significant amount. Intuitively, this is strange because there should be the exact same number of drafts each year, ranging from 1-60. The next five, MPG, PPG, FT%, FG%, and WS/48 are very similar in terms of relative importance. These features are expected, since they indicate how long and how well the players perform during each game.



Based on the drafts data in the past 20 years dataset, we visualize the relationship between categorical features and the ranking. According to this barplot “college” as one of the categorical features vs. total draft numbers, a few colleges have prominent total draft numbers, such as Kentucky and Duke having a lot higher draft numbers. It is relevant to the research question 2 as we try to understand how the bio statistics predict the draft number/round and this visualization explicitly displays the amount of drafts each different college got over the last 20 years. This motivates the research question for us to consider the relationship between the colleges and the draft numbers and suggests that college may be one of the factors that affects the drafting, which can be used in our models later on.

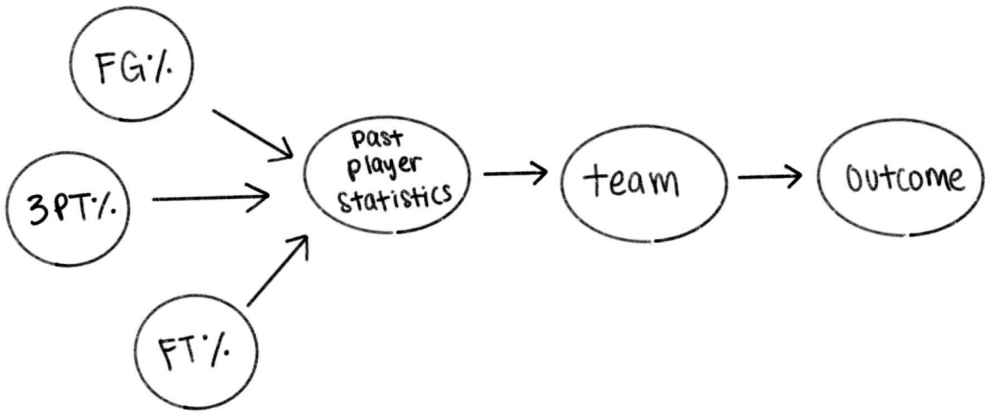
Research Question 1 Option B: Bayesian Hierarchical Modeling

Given past player statistics, can we predict the win percentage of a given team?

This question is valuable for team management and sports analysts, aiding in player acquisition strategies, team development, and predicting season outcomes. Fans can use this analysis to predict team performance after new player acquisitions or the loss of key players. We chose Bayesian Hierarchical Modeling for its ability to integrate player-specific and team-level factors, accommodating variability across players and teams, and modeling complex data relationships. This makes it suitable for sports predictive analysis, where individual performance impacts team outcomes. However, there are some limitations. The model may overfit due to many parameters and limited player data, and it's sensitive to prior choices. Its reliability decreases with rapid team changes, like trades or injuries, as it depends on historical data that may not reflect current dynamics.

Methods

Graphical Model:



Assumptions

We assume the NBA dataset from the official website is accurate and hypothesize that player statistics impact team win percentages, which drives our further research.

The groups will correspond to teams characterized by similar statistical profiles. These can be correlated with existing categories in the NBA such as divisions or conferences. From a more quantitative perspective, these teams could have similar win rates within their division.

Implementation and Statement of Results

We chose a gaussian distribution because it is suitable for modeling continuous data that tends to cluster around a mean. This could be applied to metrics like points per game, shooting percentages, etc., which usually follow a normal distribution in a population. This matches up with the data that we used as we took into account all the game metrics that have to do with the individual players. In Bayesian analysis, the Normal distribution is often used as a prior distribution because of its conjugacy properties with many likelihood functions, making the posterior easier to compute.

In [32]:
```
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
```

In [48]:
```
lin_model = LinearRegression()
X_cols = ['PLAYER_ID', 'FGM', 'FGA', 'FG_PCT', 'FG3M', 'FG3A', 'FG3_PCT',
       'FTM', 'FTA', 'FT_PCT', 'OREB', 'DREB', 'REB', 'AST', 'STL', 'BLK',
       'TO', 'PF', 'PTS', 'PLUS_MINUS', 'TEAM_CITY_Atlanta',
       'TEAM_CITY_Boston', 'TEAM_CITY_Brooklyn', 'TEAM_CITY_Charlotte',
       'TEAM_CITY_Chicago', 'TEAM_CITY_Cleveland', 'TEAM_CITY_Dallas',
       'TEAM_CITY_Denver', 'TEAM_CITY_Detroit', 'TEAM_CITY_Golden State',
       'TEAM_CITY_Houston', 'TEAM_CITY_Indiana', 'TEAM_CITY_LA',
       'TEAM_CITY_Los Angeles', 'TEAM_CITY_Memphis', 'TEAM_CITY_Miami',
       'TEAM_CITY_Milwaukee', 'TEAM_CITY_Minnesota', 'TEAM_CITY_New Orleans',
       'TEAM_CITY_New York', 'TEAM_CITY_Oklahoma City', 'TEAM_CITY_Orlando',
       'TEAM_CITY_Philadelphia', 'TEAM_CITY_Phoenix', 'TEAM_CITY_Portland',
       'TEAM_CITY_Sacramento', 'TEAM_CITY_San Antonio', 'TEAM_CITY_Toronto',
       'TEAM_CITY_Utah', 'TEAM_CITY_Washington', 'START_POSITION_C',
       'START_POSITION_F', 'START_POSITION_G']
X_train, X_test, y_train, y_test = train_test_split(game_info[X_cols], game_info['W_PC
```

In [49]:
```
lin_model.fit(X_train, y_train)
preds = lin_model.predict(X_test)
```

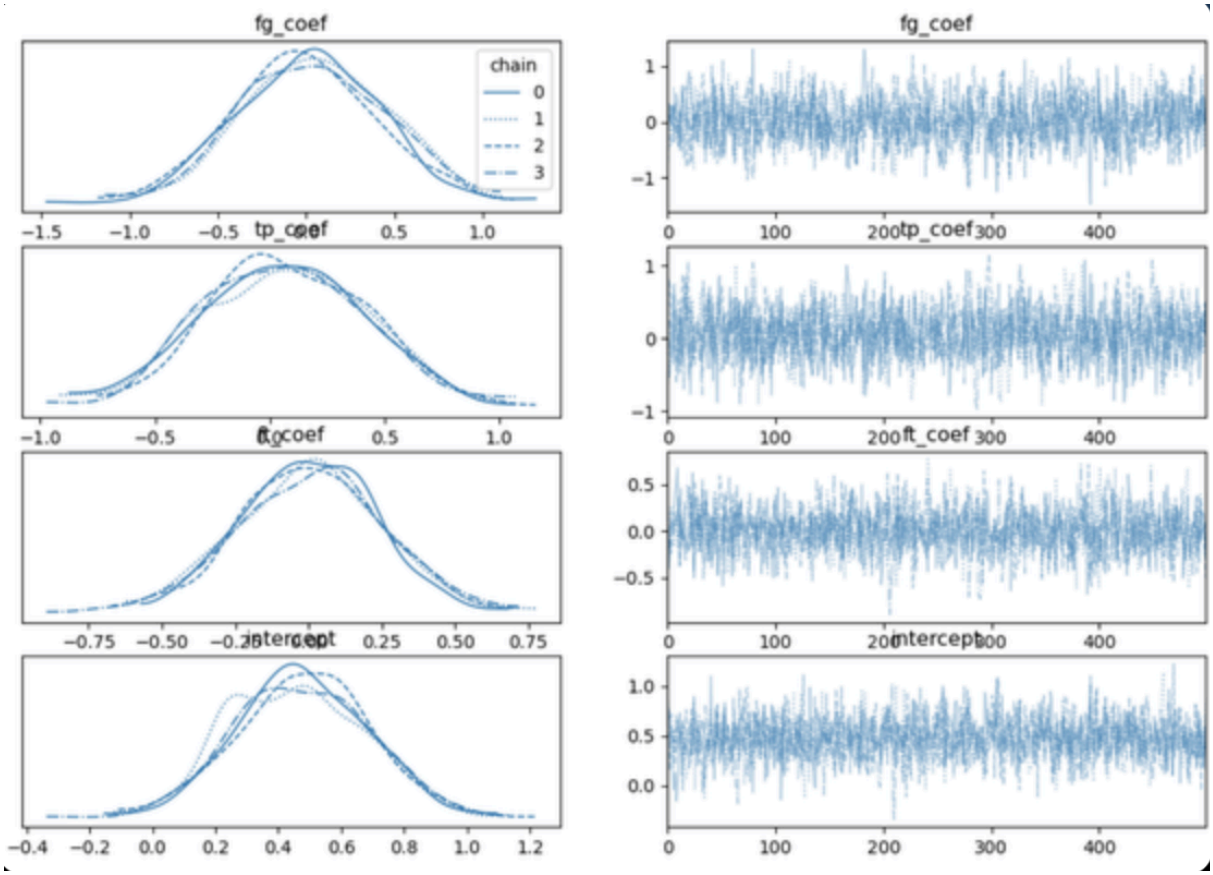
In [52]:
```
from sklearn.metrics import r2_score

# Get the accuracy score
r2_score(y_test, preds)
```

Out[52]:
```
0.21898403815343193
```

Based on our r2 value of .218, we are unable to accurately use past these player statistics to determine a team's win percentage. There were some surprises that we found during our analysis of the data, one being certain cities that a team may be from have larger coefficients than other cities, suggesting that historically certain cities that a team is from may have good win percentages, thus influencing the linear model.

Our PyMC model also suggests that we are unable to use past player statistics to determine a team's win percentage.

	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
intercept	0.475	0.211	0.075	0.863	0.006	0.004	1344.0	1297.0	NaN
fg_coef	0.013	0.412	-0.746	0.812	0.012	0.009	1249.0	996.0	NaN
tp_coef	0.064	0.350	-0.604	0.721	0.009	0.008	1570.0	1050.0	NaN
ft_coef	0.004	0.231	-0.405	0.447	0.006	0.005	1551.0	1457.0	NaN



Intercept

Mean: 0.475. This is the average predicted win percentage when all other predictors are at 0.

Standard Deviation (sd): 0.211. This indicates the variability in the intercept's estimates.

HDI_3% and HDI_97%: The values 0.075 and 0.863 define the bounds of the 94% Highest Density Interval for the intercept. This interval is the range within which the true value of the intercept is most likely to fall.

Field Goal Coefficient (fg_coef)

Mean: 0.013. This suggests a small positive effect of field goal percentage on win percentage. For each unit increase in field goal percentage, the win percentage is expected to increase by 0.013 units, holding other factors constant.

HDI_3% and HDI_97%: The values -0.746 and 0.812. The fact that this interval includes zero suggests that the effect of field goal percentage on win percentage is not statistically significant.

Three-Point Coefficient (tp_coef)
Mean: 0.064. This suggests a positive relationship between three-point percentage and win percentage.
HDI_3% and HDI_97%: The interval (-0.604 to 0.721) also includes zero, suggesting the effect of three-point percentage might not be statistically significant.

Free Throw Coefficient (ft_coef)
Mean: 0.004. This suggests a very small positive effect of free throw percentage on win percentage.
HDI_3% and HDI_97%: The interval (-0.405 to 0.447) includes zero, indicating the effect is not statistically significant.

Other Metrics
MCSE (Mean and SD): Monte Carlo Standard Error for the mean and standard deviation. Smaller values indicate more reliable estimates.
Because our MCSE mean and standard deviation were both extremely small, this means that our results are not very reliable.
ESS Bulk and ESS Tail: Effective Sample Size for the bulk and tail of the distribution. Higher values indicate more reliable estimates.
Our ESS bulk and tail estimates are as follows:
For the intercept, the ESS bulk is 1344 and the ESS tail is 1297.
For the fg_coef, the ESS bulk is 1249 and the ESS tail is 996.
For the tp_coef, the ESS bulk is 1570 and the ESS tail is 1050.
For the ft_coef, the ESS bulk is 1551 and the ESS tail is 1457.

Because almost all these values are over 1000, this indicates that our estimates are fairly precise. This means that the sampling has been effective in exploring the posterior distribution, and the MCMC chains have a low level of autocorrelation, leading to more 'independent' samples which are crucial for reliable statistical inference.

In summary:
1. Intercept: There is a moderate baseline win percentage when all predictors are at 0.
2. Field Goal Percentage (fg_coef): The effect is not statistically significant.
3. Three-Point Percentage (tp_coef): Also not statistically significant, though it shows a trend towards a positive relationship.
4. Free Throw Percentage (ft_coef): Not statistically significant with a very small effect size.

The overall results suggest that while there may be a trend towards positive relationships between these basketball statistics and win percentage, the effects are not statistically significant in our model. This could be due to high variability in the data, or it might indicate that other factors not included in the model are more influential in determining win percentage.

Uncertainty:
Intercept:
Standard Deviation: 0.211. This indicates that the estimated average win percentage when other predictors are zero could typically vary by about 0.211 units in either direction.
HDI (0.075 to 0.863): We are 94% confident that the true intercept value lies between 0.075 and 0.863. This means there's a high degree of certainty that the baseline win percentage is positive, but its exact value is quite uncertain within this range.

Field Goal Coefficient (fg_coef)
Standard Deviation: 0.412. This large deviation suggests that the impact of field goal percentage on win percentage is highly uncertain.
HDI (-0.746 to 0.812): We are 94% confident that the true effect of field goal percentage on win percentage lies between -0.746 and 0.812. The inclusion of zero in this interval indicates that we cannot be sure if the effect is positive or negative.

Three-Point Coefficient (tp_coef)
Standard Deviation: 0.350. This also indicates substantial uncertainty about the effect of three-point percentage on win percentage.
HDI (-0.604 to 0.721): We are 94% confident that the true effect lies in this range. Again, the inclusion of zero suggests uncertainty about the direction of this effect.

Free Throw Coefficient (ft_coef)
Standard Deviation: 0.231. This indicates a moderate level of uncertainty about the effect of free throw percentage on win percentage.
HDI (-0.405 to 0.447): We are 94% confident that the true effect lies between -0.405 and 0.447, including zero, which indicates that the effect could be positive, negative, or negligible.

In summary:
Intercept: We're quite confident that the baseline win percentage is positive, but there's a lot of uncertainty about its exact value.
Field Goal Percentage (fg_coef): There's a lot of uncertainty about how field goal percentage affects win percentage. It could have a positive impact, a negative impact, or no significant impact at all.
Three-Point Percentage (tp_coef): Similarly, the effect of three-point percentage on win percentage is unclear. It could be positive, negative, or negligible.
Free Throw Percentage (ft_coef): The impact of free throw percentage on win percentage is also uncertain, with the possibility of being slightly positive, slightly negative, or unimportant.

All in all, while the model suggests some trends, the uncertainty in these estimates is quite high, indicating that the precise impacts of these basketball statistics on win percentage are not clearly determined by this model.

Interpretation of results

Limitations of our methods: The Bayesian Hierarchical Model showed limitations in handling the high variability within the NBA player data. This variability may have affected the precision of our posterior estimates, as reflected in the broader Highest Density Intervals (HDIs) for some parameters.

Convergence Issues: There were challenges in convergence of the inference procedure, possibly due to the complexity of the model and the high dimensionality of the data. This was evidenced by the Monte Carlo Standard Errors (MCSEs) for mean and standard deviation, which indicated less reliable estimates.

Alternatives Tried: We experimented with different formulations of the Bayesian model. Adjustments to the prior distributions were made to improve model fit, but these changes only marginally enhanced convergence and did not significantly alter our conclusions.

Additional Data and Graphical Model Integration: Adding variables like injury history or psychological factors to the graphical model could enhance predictions. However, the subjective nature of this data presents acquisition challenges. Overall, predicting a team's win percentage in basketball, a team sport, is difficult due to the limited direct impact of individual players.

In essence, while the Bayesian Hierarchical Model provided some insights, its effectiveness was limited by the intrinsic variability in the data and the complexities of the model, suggesting a need for more refined data and potentially simpler model structures.

Research Question 2 Option C: Prediction with GLMs and nonparametric methods
Given a player's bio stats, can we predict what their draft round/number is?

Questions, algorithms, and modeling choices

The draft position of an NBA player may heavily impact their future career and earning potential, as well as affect how successful teams will be in the future. Thus, we are trying to predict what the draft number of the basketball players will be based on their past statistics, such as field goal percentage or total number of points scored. The method we will use to answer this question is prediction using GLMs and nonparametric methods. This is a good fit for our research question because the outcome that we want to predict is a rank (we assume that the lower the draft number/round, the better the player is). The dataset provides us with a number of quantitative features, which we used Random Forest Classifier's inbuilt feature selection to rank them from most to least relative importance. To analyze categorical features, we converted them into dummy variables and then performed the Random Forest Classifier.

The GLMs used were linear regression, Negative binomial, and Poisson. Linear regression was just used as a sanity check to see how it would compare with respect to the other nonparametric methods. Negative Binomial and Poisson GLMs were used because they are the best for predicting positive and discrete outcomes, which is what our rank is. Other GLMs would not be a good fit for our model, because our outcome is not binary (binomial), nor is our outcome continuous and unbounded (Gaussian).

The nonparametric methods that we will be using are decision trees and random forests. These regression methods are the best for our prediction because they can be used to predict a continuous rank (draft number). Specifically, we believe that random forests will be a better predictor because it inherently prevents overfitting, given the large number of features and players.

We evaluated each model's performance by comparing the training and test set error with that of each of the other models. In addition, for each random forest classifier we performed, we can determine their performance by their accuracy scores.

Assumptions

Some limitations of this method is that we are assuming that every team is choosing players based on the best player available to them. However, in the real world, that may not be the case. Often, teams are looking to fill a specific position, and may choose the player that is not statistically the best, but fits what they need the most. In addition, the draft also utilizes a lottery system, where “better” teams might be able to have priority in picking players.

Implementation and statement of results

```
Training set error for linear model: 12.72419314234964
Test set error for linear model:     12.236011104189418

Training set error for decision tree: 2.9548988398023552
Test set error for decision tree:     17.304387397443897

Training set error for random forest: 5.524346260752858
Test set error for random forest:     12.307820083339784
```

These are the root mean squared errors for the training and test set for linear regression, decision trees, and random forest. As we can see, the random forest model had the smallest test set error whereas the decision tree model had the largest test set error. Although the linear model did not perform the worst, the training set error was still so high that we can just assume that the linear regression model is just not a good GLM for our data set.

Random Forest Classifier:
Features: "College", "Tm", "DraftYear", "MPG", "FT%", "FG%", "PPG", "WS/48", "TOTPTS", "BPM", "TOTTRB", "TOTMP"
Accuracy: 0.029411764705882353
Features: "MPG", "FT%", "FG%", "PPG", "WS/48"
Accuracy: 0.0374331550802139
Features: 'Tm', 'Player', 'College', 'Yrs', 'G', 'TOTMP', 'TOTPTS', 'TOTTRB', 'TOTAST', 'FG%', '3P%', 'FT%', 'WS', 'WS/48', 'BPM', 'VORP', 'MPG', 'PPG', 'RPG', 'APG', 'DraftYear'
Accuracy: 0.040106951871657755

Since the random forest regressor had the lowest error rate, we decided to run random forest classifiers with the categorical variables. Surprisingly, the random forest classifier yielded intriguing results based on the selection of categorical and/or quantitative variables that we used. In the first test, the features included both categorical variables and the ten relatively most important quantitative variables. This test has the worst accuracy, of about 2.94%. The second test included only the five more important quantitative variables, because we were worried about overfitting with too many features. This test returned a higher accuracy, around 3.74%, than the previous test. Lastly, we just wanted to see what the accuracy would be if we included all the possible features, and this test gave us a 4.01% accuracy.

Negative Binomial GLM

```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:                     Pk   No. Observations:                 1868
Model:                            GLM   Df Residuals:                     1857
Model Family:        NegativeBinomial   Df Model:                           10
Link Function:                    Log   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -8076.4
Date:                Tue, 12 Dec 2023   Deviance:                       628.82
Time:                        04:06:34   Pearson chi2:                     568.
No. Iterations:                     8   Pseudo R-squ. (CS):             0.1583
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -3.2136      5.647     -0.569      0.569     -14.282       7.854
DraftYear      0.0035      0.003      1.253      0.210      -0.002       0.009
MPG           -0.0099      0.012     -0.822      0.411      -0.034       0.014
FT%            0.1407      0.158      0.893      0.372      -0.168       0.450
FG%           -0.0482      0.305     -0.158      0.874      -0.646       0.550
PPG           -0.0711      0.023     -3.027      0.002      -0.117      -0.025
WS/48          0.2160      0.731      0.295      0.768      -1.217       1.649
TOTPTS      1.312e-05   3.43e-05      0.382      0.702   -5.41e-05    8.04e-05
BPM            0.0069      0.016      0.444      0.657      -0.024       0.038
TOTTRB     -8.792e-05   2.83e-05     -3.109      0.002      -0.000   -3.25e-05
TOTMP       6.067e-06   1.68e-05      0.360      0.719   -2.69e-05    3.91e-05
==============================================================================

Confidence Intervals:
                   0         1
const     -14.281745  7.854479
DraftYear  -0.001991  0.009052
MPG        -0.033557  0.013718
FT%        -0.168053  0.449503
FG%        -0.646145  0.549745
PPG        -0.117100 -0.025060
WS/48      -1.217117  1.649034
TOTPTS     -0.000054  0.000080
BPM        -0.023666  0.037528
TOTTRB     -0.000143 -0.000033
TOTMP      -0.000027  0.000039
```

For every feature with 95% confidence:

- DraftYear: For each additional unit increase in 'DraftYear,' the ranking is expected to change between approximately e^-0.00199 and e^0.00905.
- MPG: With a one-unit increase in 'MPG,' the ranking changes between approximately e^-0.03356 and e^0.01372.
- FT%: A one-unit change in FT% is associated with a change in ranking between approximately e^-0.16805 and e^0.44950.

- FG%: A one-unit change 'FG%' within the interval results in a change in ranking from approximately e^-0.64615 to e^0.54975.
- PPG: With every one-unit increase in 'PPG,the ranking decreases between approximately e^-0.11710 and e^-0.02506.
- WS/48: The ranking is expected to change with one unit change in 'WS/48' within the interval from approximately e^-1.21712 to e^1.64903.
- TOTPTS:For every one-unit increase in 'TOTPTS,' the ranking is estimated to change between approximately e^-0.000054 and e^0.000080.
- BPM: A one-unit change in 'BPM' is associated with a ranking change between approximately e^-0.02367 and e^0.03753.
- TOTTRB: With each one-unit increase in 'TOTTRB,' the ranking is predicted to decrease between approximately e^-0.000143 and e^-0.000033.
- TOTMP: A one-unit change in 'TOTMP' causes changes in ranking within the interval from approximately e^-0.000027 to e^0.000039 with 95% confidence.

Poisson GLM

```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:                     Pk   No. Observations:                 1868
Model:                            GLM   Df Residuals:                     1857
Model Family:                 Poisson   Df Model:                           10
Link Function:                    Log   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -10696.
Date:                Tue, 12 Dec 2023   Deviance:                       12107.
Time:                        04:06:34   Pearson chi2:                 1.21e+04
No. Iterations:                     5   Pseudo R-squ. (CS):             0.9897
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -3.2873      0.986     -3.333      0.001      -5.220      -1.354
DraftYear      0.0036      0.000      7.257      0.000       0.003       0.005
MPG           -0.0083      0.002     -3.522      0.000      -0.013      -0.004
FT%            0.0737      0.026      2.797      0.005       0.022       0.125
FG%           -0.0319      0.049     -0.651      0.515      -0.128       0.064
PPG           -0.0676      0.005    -13.190      0.000      -0.078      -0.058
WS/48          0.0558      0.115      0.484      0.628      -0.170       0.282
TOTPTS      2.376e-05    8.4e-06      2.829      0.005     7.3e-06    4.02e-05
BPM            0.0069      0.002      2.773      0.006       0.002       0.012
TOTTRB     -8.978e-05   7.11e-06    -12.619      0.000      -0.000   -7.58e-05
TOTMP       1.112e-06   3.89e-06      0.286      0.775   -6.52e-06    8.74e-06
==============================================================================

Confidence Intervals:
                  0         1
const     -5.220385 -1.354230
DraftYear  0.002606  0.004534
MPG       -0.012977 -0.003697
FT%        0.022055  0.125355
FG%       -0.127808  0.064090
PPG       -0.077697 -0.057594
WS/48     -0.170268  0.281896
TOTPTS     0.000007  0.000040
BPM        0.002024  0.011783
TOTTRB    -0.000104 -0.000076
TOTMP     -0.000007  0.000009
```

For every feature with 95% confidence,

- DraftYear: For each additional unit increase in 'DraftYear,' we predict the ranking to increase by between approximately e^(0.0026) and e^(0.0045).
- MPG: With a one-unit increase in 'MPG,' we predict the ranking to decrease by between approximately e^(-0.01298) and e^(-0.00370).
- FT%: For each one-unit increase in 'FT%,' the ranking increases in between e^(0.02206) and e^(0.12536).

- FG%: A one-unit change in 'FG%' will change the ranking by between approximately e^(-0.12781) and e^(0.06409).
- PPG: With every one-unit increase in 'PPG,' the ranking decreases by a factor between approximately e^(-0.07770) and e^(-0.05759).
- WS/48: The ranking changes when one unit changes in 'WS/48' within between around e^(-0.17027) and e^(0.28190).
- TOTPTS: For each one-unit increase in 'TOTPTS,' the ranking increases in between approximately e^(0.000007) and e^(0.000040).
- BPM: A one-unit change in 'BPM' is associated with a change between approximately e^(0.00202) and e^(0.01178).
- TOTTRB: With each one-unit increase in 'TOTTRB,' the ranking is predicted to decrease by a factor between approximately e^(-0.000104) and e^(-0.000076).
- TOTMP: The ranking is expected to vary with changes in 'TOTMP' within a factor between approximately e^(-0.000007) and e^(0.000009).

Interpretation of results

The Negative Binomial GLM performed better, and we can tell by looking at the log-likelihoods. For the Negative Binomial GLM, it had a log likelihood of -8076.4, whereas the Poisson GLM had a log likelihood of -10696. The Negative Binomial GLM also adhered to the data more closely, because it had a lower deviance of 628.82, compared to the Poisson GLM's deviance of 12107. I think overall, the Negative Binomial GLM is a viable model to use for similar datasets and predictions in the future.

For the nonparametric test, the random forest classifier performed better. However, because the accuracy is still so low (less than 5% for all tests), I would not be confident at all in applying this method to future datasets.

Since we are implementing our GLMs using the entire dataset of drafted NBA players, we are only using the Bayesian implementation. This means that we are only aiming to find the unfixed parameters of the distribution that generated our data.

For GLM, the limitation of negative binomial is that it is more complex than Poisson, and can cause increased complexity of model and interpretation. The limitation of Poisson is that it assumes the mean and variance of the response variable are equal; however, in real-world scenarios like for NBA datasets, it may cause overdispersion where the variance is greater which would make the Poisson model be a less accurate fit. Poisson is also sensitive to outliers.

For nonparametric models, the disadvantage of decision trees is overfitting and being largely dependent on training data. The disadvantage of random forests is that we cannot describe the relationships within the data, and requires more storage and computational power to get accurate results.

Some additional data that would be useful for improving our model would be having additional features on the players such as height, weight, and age. That is because those inherent characteristics such as height and weight are very important indicators to how well NBA players can play, and the NBA draft historically also favors younger players. Other data that could be useful in creating a more accurate model is knowing the order of the teams picking players.

According to the GLM model summary generated, the uncertainty of the results can be determined. The smaller/narrower confidence interval range indicates more confidences and precise estimates, whereas a larger/wider confidence interval range indicates greater uncertainty. As per the results, for Poisson GLM, Features such as TOPMP, DraftYear, TOTTRB, and TOTMP have a smaller CI compared to other features. The larger confidence interval range would be features like TOTPTS, FT%, FG%, WS/48, BPM, PPG, and MPG, and these features indicate greater uncertainty in estimates. For Negative Binomial GLM, Features such as TOTTRB, TOTPTS, DraftYear, BPM have a smaller CI. FG% has a relatively wide CI; furthermore, features like WS/48, FT%, PPG, and MPG all have a moderately wider CI.

Both GLM are mostly consistent with wider and narrower confidence intervals features. Overall, FG% (field goal percentage), FT% (field goal percentage), PPG (Points per game) , MPG (Minutes Played Per Game), and WS/48 (win share per 48 minutes) are the features that showed the most uncertainty in the estimates. This may be due to the complexity of the model where these features lead to increased uncertainty. Furthermore, the dataset may contain a significant amount of noise or unexplained variability which can contribute to the higher uncertainty in the estimates.

Conclusions

Research Question 1: "Given past player statistics, can we predict the win percentage of a given NBA team?"

Overall, the Bayesian Hierarchical Model indicated limited predictive power of player statistics on team win percentage. Results are moderately generalizable within the NBA context, but may not extend to other sports or lower levels of competition, where variability makes it more difficult to generalize. Our results show that teams could focus more on team dynamics and less on individual player statistics for strategy development because it seems like an individual player does not do much to impact a team's win percentage directly. High data variability and model complexity were limitations. Future research could include more dynamic variables like player trades or health status, as well as how well certain players work together with others.

Research Question 2: "Given a player's bio stats, can we predict what their draft round/number is?"

Both GLMs and nonparametric methods like random forests were used. Overall, the models showed low predictive accuracy. The generalizability is very limited due to the specific nature of NBA draft processes, and how they change over time. Our group suggests caution in solely relying on statistical models for draft decisions; incorporating qualitative assessments may be beneficial, as they do not seem to correlate much. Our group also encountered some limitations, such as the inability to include certain key biostatistics like height or weight. Future studies could incorporate these variables for more robust predictions.