NCAA Men's Basketball Analysis

Qianhan Zhang, Jiannan Zhao, Wangqian Ju, Haoran Zhang, Peifen Lyu

Abstract:

In the report, we are interested in analyzing National Collegiate Athletic Association (NCAA) Men's Basketball dataset on two levels: individual-player level and player-player level.

For the individual level, we mainly did exploratory data analysis and used ANOHT and other techniques to investigate associations between features. Then we used these associations and information to answer questions we are interested in. For example, (a) how to reasonably measure the performance of an individual player, and (b) which features mainly affect the number of points a player can get.

For the player-player level, we wonder (a) how to effectively measure the communication between players in a team, and (b) what is the teamwork effect in NCAA games. We focused on three teams: the 2016 Champion, Villanova University, top 8 team, Oklahoma, and UC Davis. We draw a directed graph for each team's players assists in games: each node is a player and an arrow will be drawn from player A to player B if A assists B to shoot a basket at some game in 2016 season. We also draw dots on a basketball field to indicate where the player stands when he scored with/without teammate's assists to compare the difference. After comparison, we find that there are more players who assists each other in Villanova team than in UC Davis team. We also find in both teams that three-point field goals tend to be more of a teamwork and two-point shot rely more on player's individual act.

1. Introduction:

Nowadays, we are able to record, collect, and analyze the information generated in different sports games. The statistics inside sports has become an important tool for coaches and the public to evaluate the performance of the teams and players. Numerous methods of statistical analysis and models have been created with the attempt to precisely and objectively evaluate the performance of particular players and whole teams. Combined with the quality and quantity of information available on the Internet, the new approaches that have been developed for analyzing the game can help us to have a deeper understanding of the sports games.

Due to its popularity, basketball game has become one of the most analyzed sports disciplines. In this project, we are interested in analyzing National Collegiate Athletic Association (NCAA)

Men's Basketball, one of the famous annual sporting events in the United States, especially how to effectively evaluate the performance of the players and teams based on different features and levels. The dataset is introduced in Kaggle. It contains information about NCAA Basketball games, teams, and players. Game data covers play-by-play and box scores back to 2009, as well as final scores back to 1996.

This extensive dataset enables us to explore and analyze the performance based on two levels: individual-player level and player-player level:
For the individual level, we are interested in the questions like (a) how to reasonably measure the performance of an individual player, and (b) which features mainly affect the number of points a player can get.
For the player-player level, we wonder (a) how to effectively measure the communication between players in a team, and (b) what is the teamwork effect in NCAA games.

To answer those questions, we investigate the associations between different features (weight, height, position, etc) based on the mutual conditional entropy, use analysis of Histogram (ANOHT) visualize the associations, build network graph for the player-player interaction, model and visualize the connections in the court and other techniques. In the end, we combine the statistical analysis and information for both levels to figure out the answers.

With this information and analysis, we are able to establish a reasonable statistical report and have a better understanding of what is happening on the court. This statistical report might allow people to evaluate the technical and tactical efficiency of individual players and teams, and compare them during a single game or even the whole season. It is meaningful for the team and coaches since they can adjust their training and strategies accordingly. The information and analysis are also beneficial for the public and players since it can be used to quantify the performance and indicate how to improve. We hope that our project can serve as a demonstration of the value of the data and analysis in resolving real-life problems.

2. Material

The dataset is introduced on Kaggle, and we then traced back to the original dataset on Google BigQuery.(https://console.cloud.google.com/bigquery?project=arctic-defender-204523&folder&organizationId=558550560619&p=bigquery-public-data&d=ncaa_basketball&page=dataset) We downloaded the dataset locally for data analysis. There are 10 tables in this dataset, and 8 of them are useful to us:

1. Mbb_games_sr.csv: Team-level box scores from every men's basketball game from the 2013-14 season to the 2017-18 season. Each row shows both teams' stats for that one game.
2. mbb_historical_team_games.csv:Final scores for men's basketball games, starting with the 1996-97 season. Each game is included twice, with one entry per team.
3. mbb_historical_teams_seasons.csv:Season record information for Men's Basketball, starting with the 1894-95 season. Each game is included twice, with one entry per team.
4. mbb_historical_tournament_games.csv:Game score information from Men's Basketball games, starting with the 1984-85 tournament. Each row shows one game.
5. mbb_pbp_sr.csv:Play-by-play information from men's basketball games, starting with the 2013-14 season. Each row shows a single event in a game.
6. mbb_players_games_sr.csv:Player-level box scores from every men's basketball game from the 2013-14 season to the 2017-18 season. Each row shows a single player's stats in one game.
7. mbb_teams.csv:General information about the 351 current men's D1 basketball teams.
8. Mbb_teams_games_sr.csv:Team-level box scores from every men's basketball game from the 2013-14 season to the 2017-18 season. Each row shows a single team's stats in one game. This data is identical to mbb_games_sr, but is organized differently to make it easier to calculate a single team's statistics

3. Method:

3.1. Individual-player level:

To effectively analyze the performance at the individual-player level, we decide to analyze the data on different features and their associations since it might be inconclusive to evaluate a player’s performance only based on the points they can get at each game; there are also more criteria needed to be considered. In this report, the statistical models, Analysis of Histogram (ANOHT) and mutual conditional entropy, are used to analyze the data.

ANOHT provides a way of converting continuous features into categorical ones. As mentioned in Dr. Hsieh’s paper[citation], the possibly gapped histogram generated from the ANOHT allows us to re-normalize numerical features into digital-categoricals in a reasonable way. The re-normalization from real-valued into digital-categorical then leads to the application of combinatorial information theory and thus allows us to compute the mutual conditional entropy to measure the associations among all features that we are interested in. Additionally, it is a good tool for us to explore the data. With color-coded and possibly gapped histogram, we are able to investigate the directed association from one feature to another feature.

Mutual information measures the dependence and association between two variables. Among the models in information theory, mutual conditional entropy quantifies the amount of information needed to describe the outcome of a random variable Y given that the value of another random variable X. In this report, we are following the equation from Dr. Hsieh's paper[citation], stated by Eqn.1:

$$\frac{H(Y|X)}{H(Y)} =- \sum p(x,y) log \frac{p(x,y)}{p(x)} \ (Eqn.1)$$

where H(Y) is the entropy of variable y, calculated by Eqn. 2:

$$H(Y) = \sum p(y) log[p(y)] \ (Eqn.2)$$

The conditional entropy H(Y|X) gives a real-valued number from 0 to 1. The value of 0 indicates the strongest association and that the value of Y is completely determined by X. Conversely, the value of 1 indicates the independence of the two variables. In other words, if two features have high association, the mutual conditional entropy value will be low. This understanding of the associations is important and meaningful because it reveals the dependency between different features and their importance, it will also offer us insights in further analysis or even modeling.

Our primary interest is that how to evaluate the performance of NCAA players based on their gameplay data. Given the large number of features provided by the dataset, twenty-one features we are interested in and contain more useful information, including player's personal information (weight, height, starter, etc) and gameplay related skills (assist, rebonds, three-points count, etc). One thing worth noting is that, since the original sample size which contains 115072 observations, is large, we subset our dataset and firstly focus on the games in season 2017 - 2018 and teams which were the top eight or the last eight.

3.2. Player-Player level:

We extract the names and points scored through assists of Players from the columns named "event_description" and "points_scored" by regular expression. For the column of event_description, it has fixed format that Player A makes two or three points shot (Player B assist). For the column of points_scored, it only contains the points got by assists but half of them are missing.

Then we are going to draw the network graph by the package named networkx. We only use the names of assisters and player assisted as nodes and label them by names, and the assists between them as edges with arrows.

Since we have all the coordinates of the assist positions, we are going to visualize the distributions of these positions in the real basketball court. We use the mathplotlib library to plot the basketball court in real size, and then scatter all the positions of points scored without/with assist.

4. Graphic and Findings:

4.1 Individual-player level:

To analyze the performance of players at individual level, we decide to study the associations between different features, and then figure out how those important features affect players' gameplay related skills. The ANOHT is used first to convert numerical variables into categorical ones. The resulted groups are used to calculate the mutual conditional entropy to show the associations and dependencies between two features. The gapped histogram from ANOHT is then used to visualize the detailed relation inside the pairs of features which have strong associations. We are also interested in how those feature selections will change according to different positions in the game, rankings of the teams in the year and year by year. The statistical data of each player at each game from the Top 8 teams in the 2017-2018 year serve as the base to compare with other data.

4.1.1 ANOHT

ANOHT allows us to generate colored and possibly gapped histogram as mentioned in section 3.1. It firstly serves as a way of converting continuous variables into categorical ones. To do the conversion, we are following the method suggested in Dr. Hsieh's paper[citation]. We discovered a bug in the source code in the 'GappedHist' package. The bug lies between line 23 to line 30. In the source code, when the height of the dendrogram of the first tree is equal to that of the second tree. The local variable 'selected' and 'not.selected' will not be initialized. Leading the function to crash. We fix it by randomly choosing a set of tree, the result of clustering is the same.

Figure 1 is an example of a result of ANOHT. It is the gapped histogram of the feature "point" and is color-coded by the feature "starter".

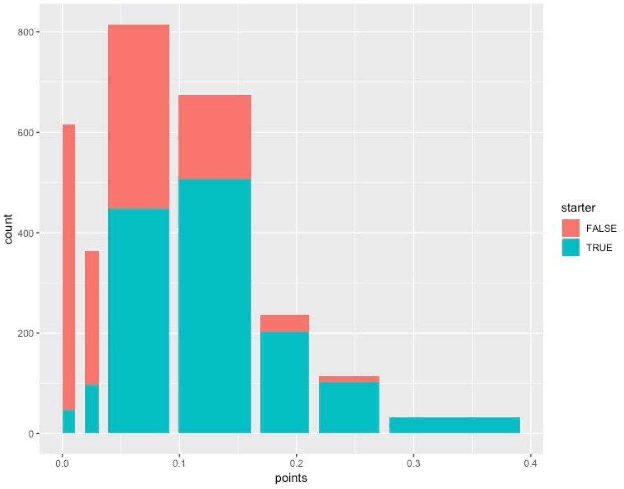


Figure 1. The histogram of points colored by starters for players in Top 8 teams in 2017

It shows that ANOHT converts the continuous variable "point" into a categorical one with 7 levels, calculated by histByDess() based on hierarchical clustering. It indicates the association between these two features: For the highest level of feature "point", there are only starter players; for the lowest level, the portion of non-starter players is much larger than that of starter players. This also suggests that starter players have more better ability to score, which makes much sense to us since starter players are usually the top players of a team in basketball.

4.1.2 Mutual Conditional Entropy and Heatmap

After using ANOHT to convert convert all the interested numerical features into categorical ones, we are able to calculate the mutual conditional entropy of any two of all the features, introduced in section 3.1. The mutual conditional entropy of all pairs of features is then converted into a heatmap (Figure 2) to visualize the result.

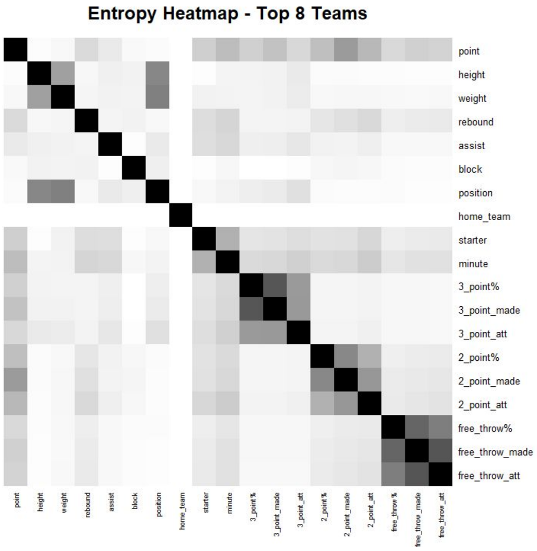


Figure 2. Mutual entropy heatmap of different features for the players in the top 8 teams in 2017-2018

In this heatmap, black indicated mutual conditional entropy of 0, the strongest association, while white indicates mutual conditional entropy of 1, which means the two features are independent. The heatmap gives us a lot of information about the data. For example, the dark grey color between the feature “points” and other features, “starter”, “minute”, “2_points_made”, “3_points_made” and “free_throw_made” shows that a player’s ability to score is closely related to whether or not he is a starter player, the time he plays, the number of field goals he makes and attempts, and etc. It’s also clear that the feature “home_team”, which indicates whether the gameplay is in the player’s home court, has almost no association with all other features. This also indicates that the players of the top 8 teams have consistent performance regardless of they are in home court or not.

We can also draw dendrogram for all the features based on the heatmap (Figure 3).

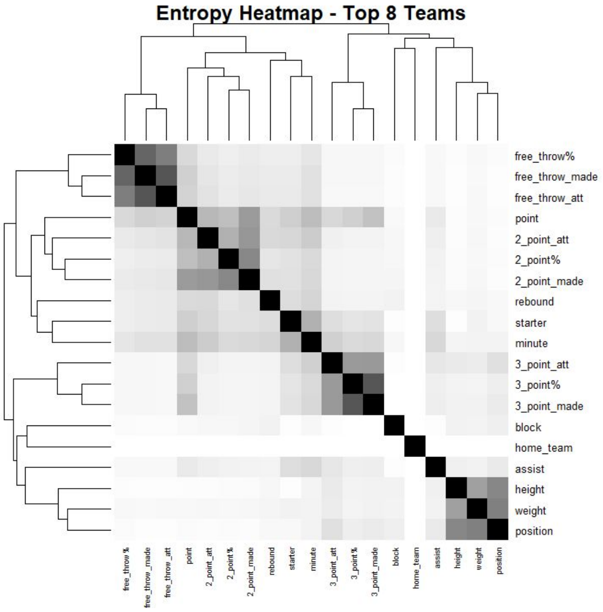


Figure 3. Dengrogram of different features for the players in the top 8 teams in 2017-2018

The block structure can be seen in this heatmap with dendrogram. The block structure indicates features that are close to each other. Take the most interested feature “point” as an example, this plot shows that it has high association with two-point shots, then number of rebound, whether or not the player is a starter, his play time, then free-throw shots. Normally, three-point shots is considered as an important gameplay skill in basketball. The data shows that it has high association with feature “point”, but it has high association with the player’s height, weight, and position, as well.

The results mentioned above suggest that the combination of ANHOT (section 4.1.1) and mutual conditional entropy (section 4.1.2) can provide reliable statistical models to study the associations between different features and help us to select the critical pairs based on different criteria.

4.1.3 Feature Selection based on Positions

A player’s position indicates his role in his team. A guard player might focus on making shots, while a center player might focus on other skills, such as block or rebound. Thus, to measure the performance of a player, we should not only consider the points he makes, but also other features based on the player’s position or role in the team. The feature “position” might also introduce heterogeneity into our analysis, compared to the one which combine all the players’ information (Figure 2). Therefore, we separated the data based on the player’s position, and did ANOHT and calculated the mutual conditional entropy.

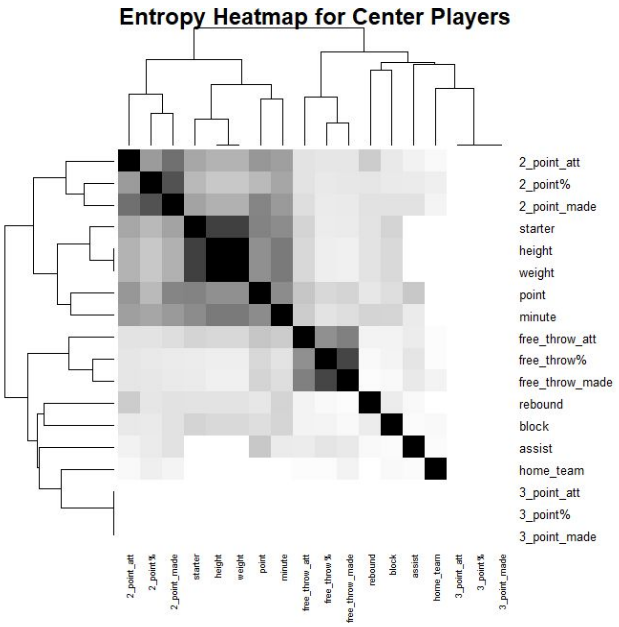


Figure 4. Dengrogram of different features for the Center players in the top 8 teams in 2017-2018

The heatmap of Center players (Figure 4) shows that center players’ ability to score is closely related to their play time, whether they are starter, and ability to shoot two-point shots, which corresponds to the results gained from the analysis of players at all positions above (Figure 3). However, the center player’s ability to score is also related to their height, weight, and ability to assist. To our surprise, even though the feature “block” is closer to the feature “rebound”, the center players’ ability to score has higher association with their ability to assist, compared with their ability to block and to rebound. This might encourage us to draw a gapped histogram of feature “point” and color-coded by the feature “assist”. Another thing worth noting is that, the center players in our dataset didn’t make three-point shots, thus we have the white band in the heatmap. It indicates that the ability to make three-point shots are not crucial for center players as to other positions.

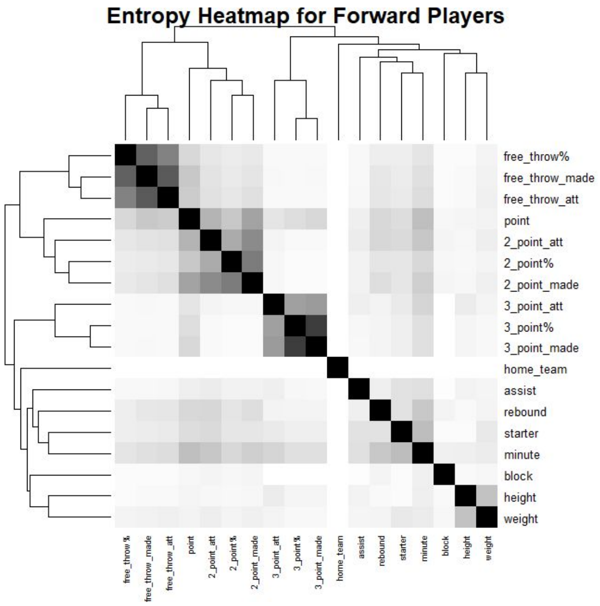


Figure 5. Dengrogram of different features for the Forward players in the top 8 teams in 2017-2018

The heatmap of forward players (Figure 5) shows that their ability to score is closely related to their ability to shoot two-point shots and free-throw shots, then three-point shots, which matches with the results gained from the analysis of players at all positions above (Figure 3). In contrast with other positions, rebound matters more to forward players, but their height and weight don't have high association with their ability to score

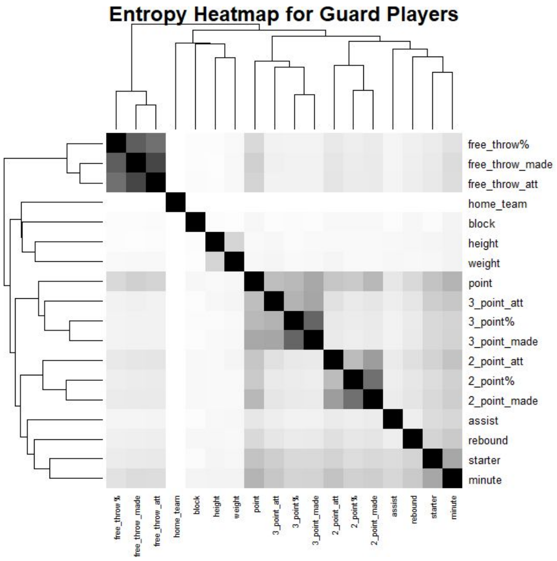


Figure 6. Dengrogram of different features for the Guard players in the top 8 teams in 2017-2018

The heatmap of guard players (Figure 6) also shows a coincidence with the results gained from the analysis of players at all positions above (Figure 3). However, compared to other positions, where two-point shots are more important than three-point shot, the three-point ability is more closely related to the player's ability to score. Also, "rebound" and "assist" is part of the block structure of "point", while "block", "height", and "weight" are outside of the block structure.

4.1.4 Feature Selection based on Team Rankings

It is always interesting to investigate why some team win and why some team lose. Therefore, a heatmap that represents the difference of association between high ranking team and low ranking team would be useful to indicate the aspects that a team should focus on if they want to perform better.

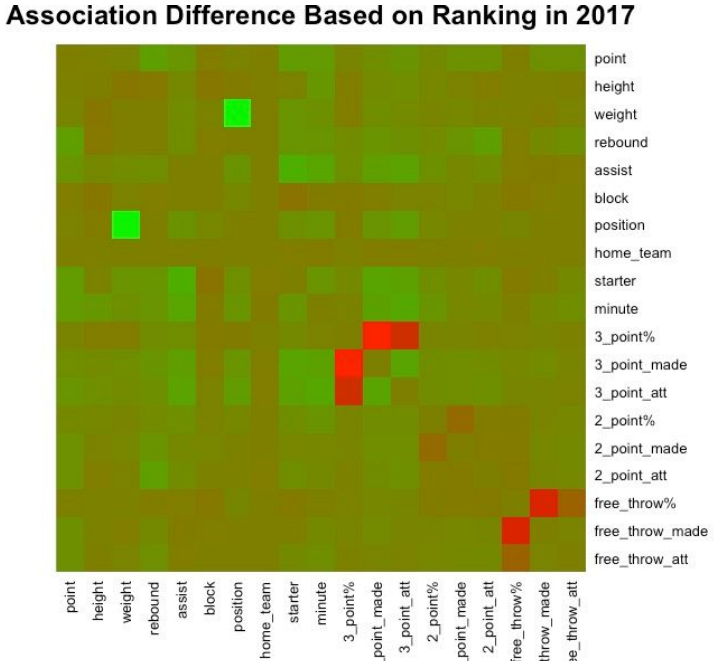
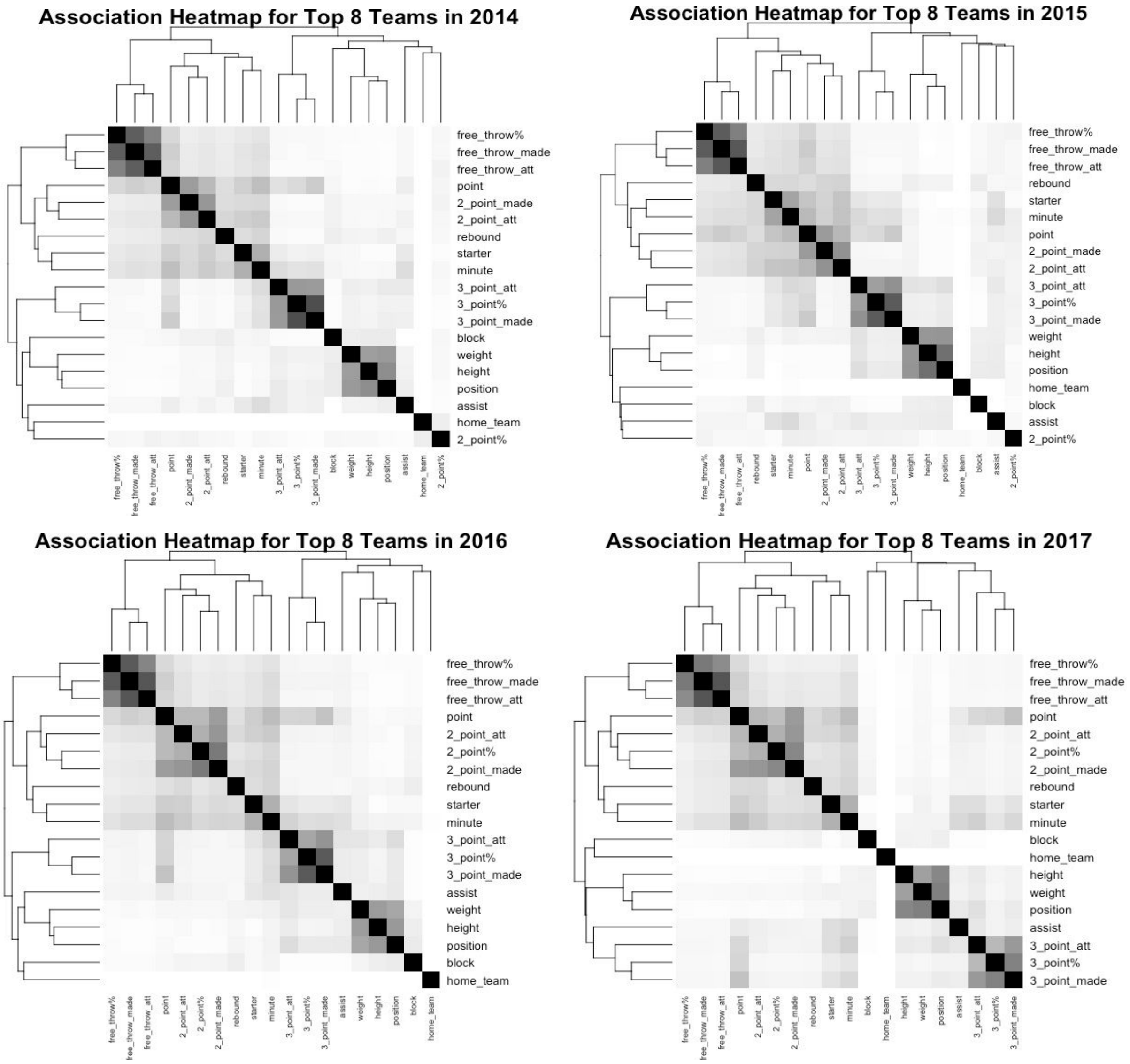


Figure X. Heatmap of the difference in association between top 8 teams and last 8 teams in 2017

From Figure X, we can see the difference of feature associations between top 8 teams and last 8 teams in 2017. Green means that the top 8 teams has a greater feature association, and red means that last 8 teams has a greater association. It seems that the better performing team has a higher association between weight and position. It suggests that the coach should assign each player's position based more on weights. Additionally, better performing team has higher associations between starters and their ability to assist, which suggests that cooperation within the team is another crucial factor to win. Moreover, It seems that worse performing team has a stronger association between 3 points and 3 point percentage, free throw made and free throw percentage.

4.1.5 Feature Selection based on Years

Another interesting aspect is to investigate on the trend of high ranking teams performance over the years so that we can have a better understanding of how NCAA game changes and the focus on important features shifts. It should have its own characteristics, and we can use those characteristics to help coaches and players understand how to play better in NCAA at different times.



Association Heatmap for Top 8 Teams in 2014

Association Heatmap for Top 8 Teams in 2015

Association Heatmap for Top 8 Teams in 2016

Association Heatmap for Top 8 Teams in 2017

Figure X Association Heatmap for Top 8 teams from 2014 to 2017

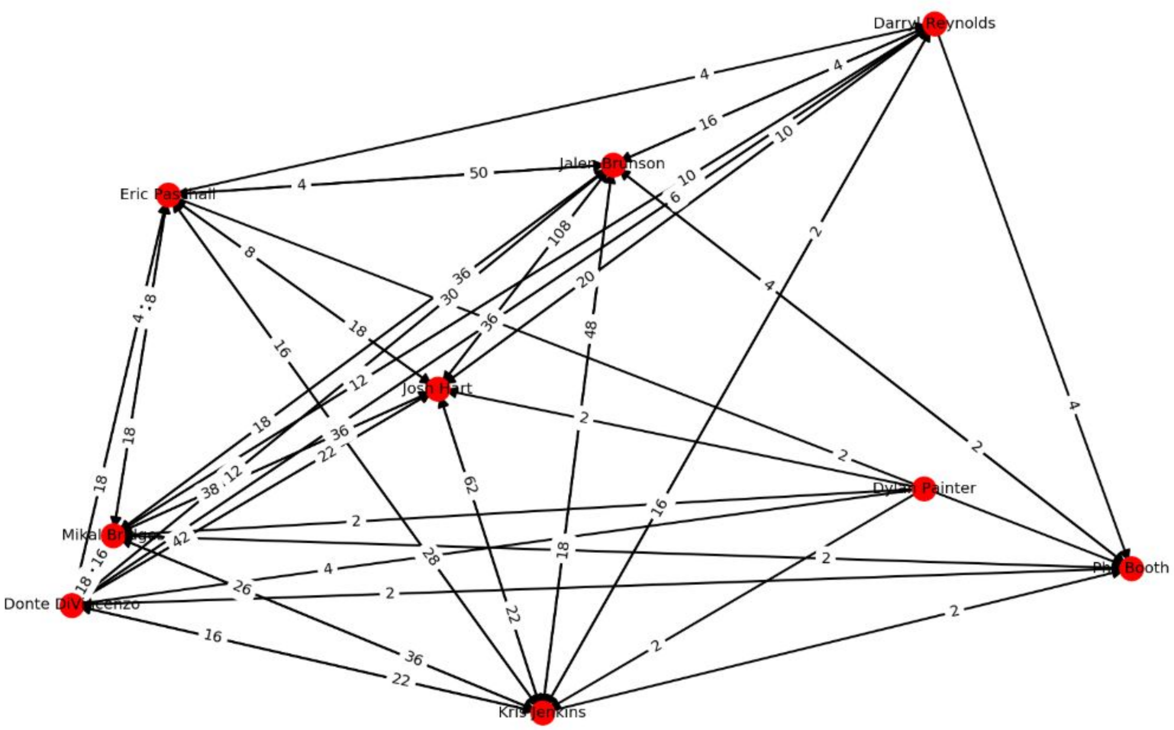
From the figures above, we can analyze the trend of associations between different features for high ranking teams over the years. Based on the block structure, we can see that the trend is pretty stable that the player's ability to score is more closely related to whether or not he is a starter player, the time he plays, the number of field goals he makes and attempts, and etc.

The player's ability to score is more independent on players' ability on assist, weight and height. It suggests that NCAA tournament game relies more on individual player's ability rather than team work, and the main may to score high points is more on free throws and field goals, rather than 3 points which is very popular recently in NBA games. It makes sense since the team player changes rather frequently in NCAA, so a team has very little time to truly master a set of strategies. Therefore, individual player's ability plays a more important role than team work does.

In addition to player's ability to score, we also take a look at how different features affect the starter selection. In 2014, the ability to make two-point shots are more closely related to the starter selection; however, the ability to make three-point shots become more important as time past. In 2017, it shows stronger association with the starter selection compares to two-points shots. It suggests that NCAA tournament game give players who can make more three-point shots more opportunities as starters. The time the player can play at each game also follows similar trend.

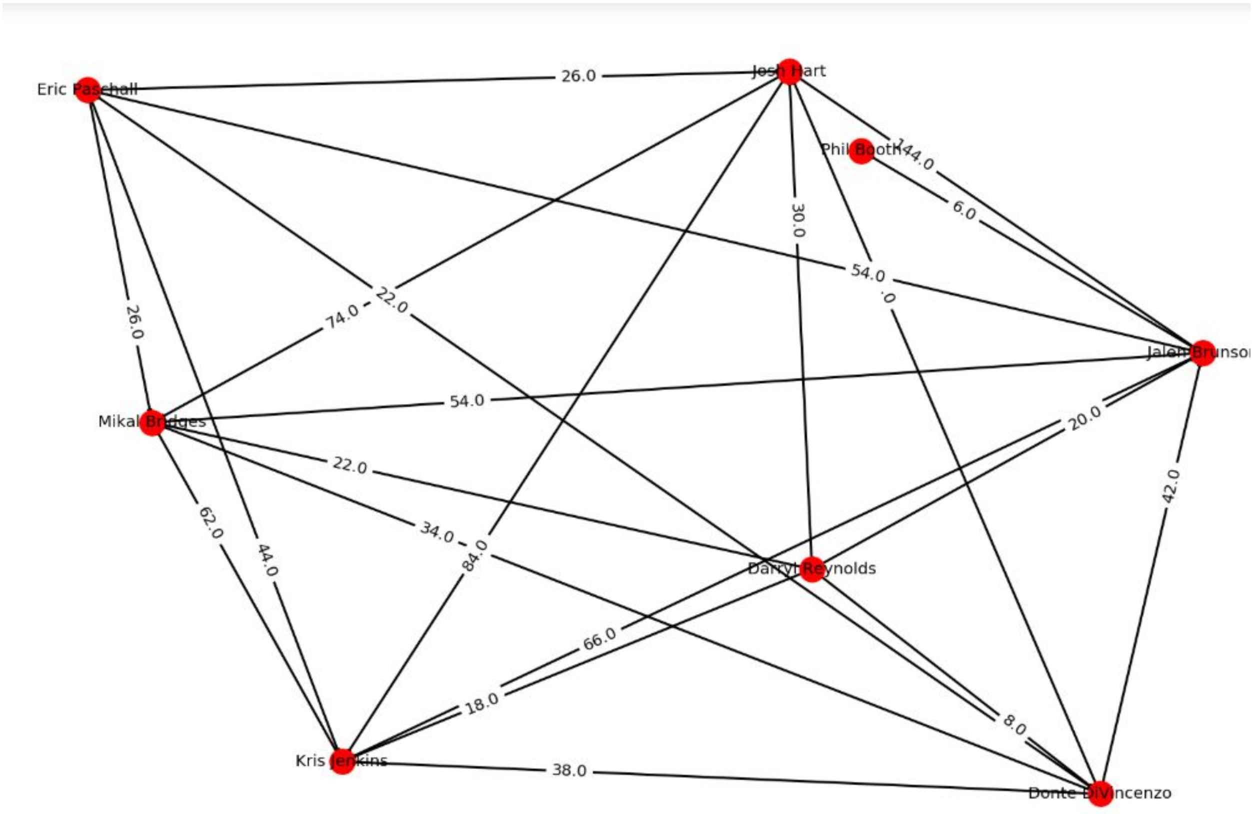
4.1. Player-Player level:

4.2.1 Player Assists Network



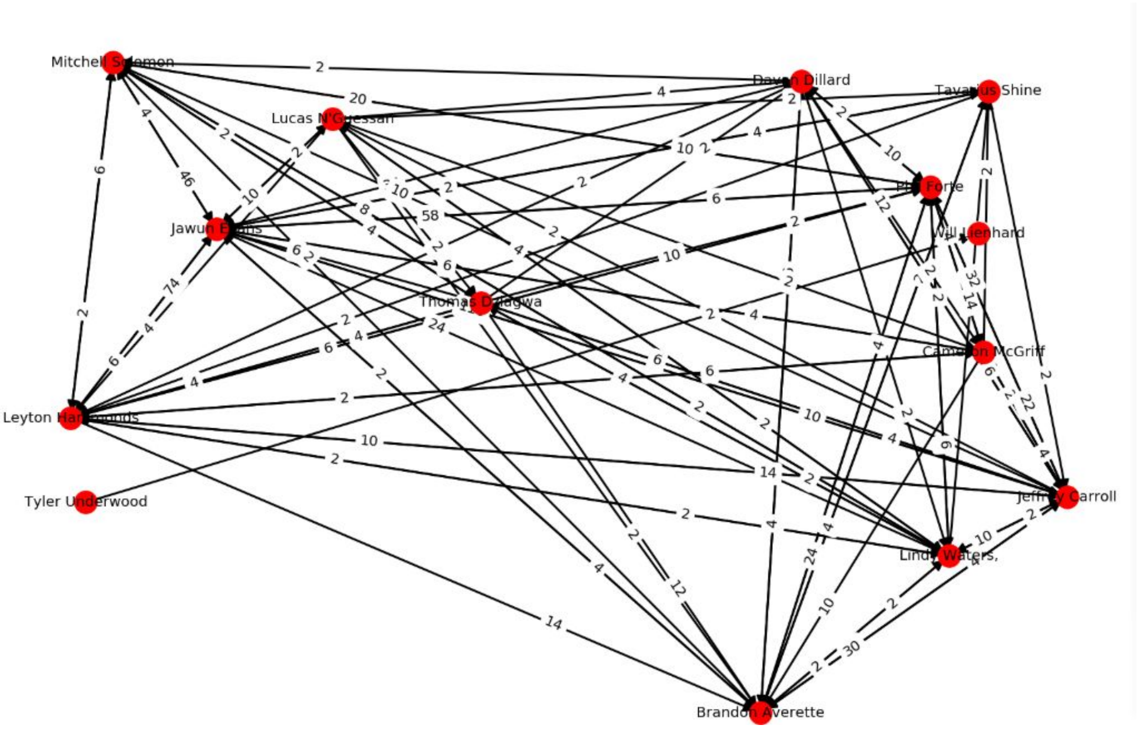
Specific Assists between Players in Villanova

For the figure of Villanova, the winner of season 2016, it contains all the connections between different pairs of players. Arrows means the assist from player A to player B. For example, the edge between Darryl Reynolds and Kris Jenkins includes two numbers, 2 and 16. Two means Darryl had been assisted by Kris for two times and Kris had been assisted by Darryl for sixteen times. We can find the largest time of assists is 108 from player named Josh Hart to player called Jalen Brunson. What's more, based on the graph, we can find Eric Paschall, Jalen Brunson, Kris Jenkins and Mikal Bridges are core players of this team since they all got assists from other 7 different players. We can roughly conclude that the connection between these players is best in this team and Jalen Brunson may be the best player in this team.



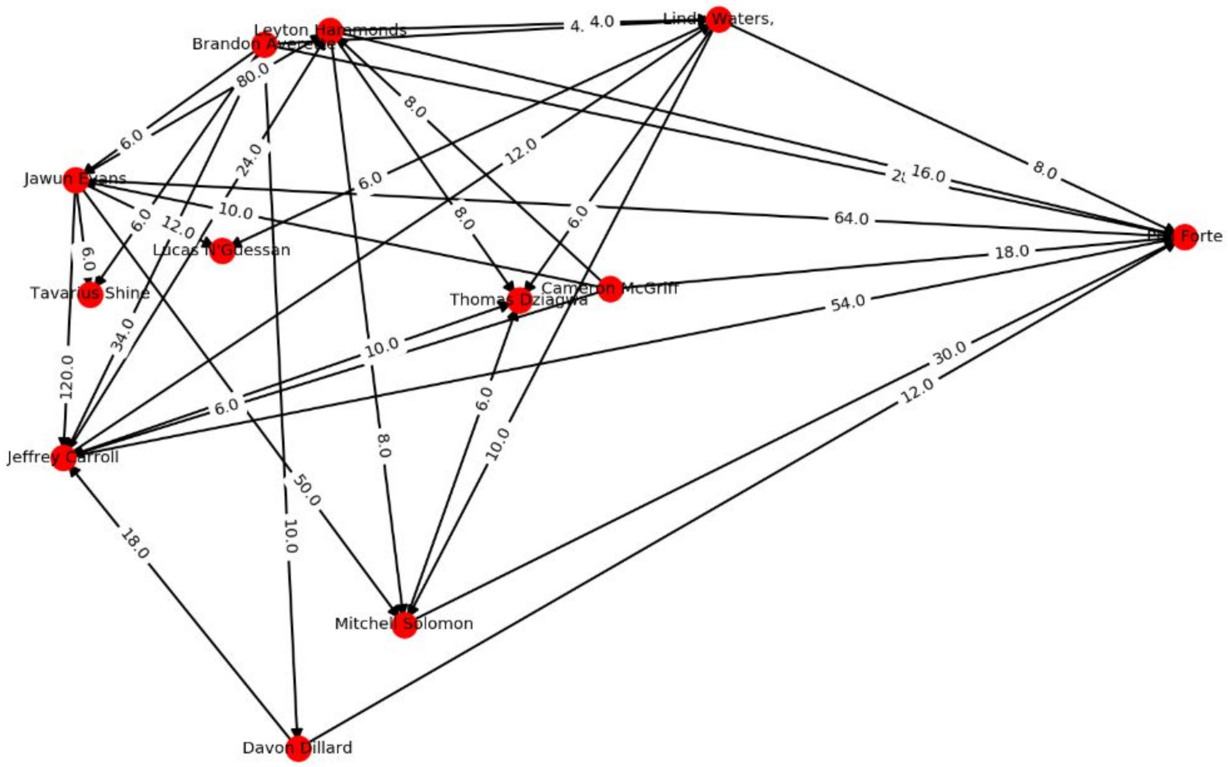
Total Assists between Pairs of Players in Villanova

From the figure above, the edges between different pair of players are the total assists among them. For example, the edge between Eric Paschall and Kris Jenkins is 44 which means there are total 44 assists. Among all the edges, the largest number is 144, between player called Jalen Brunson and Josh Hart. Now, we can conclude confirmly that these two players are best partners in this team.



Specific Assists between Players in Oklahoma

We select Oklahoma, which was a top 8 team in the season 2016, because it has nearly as the same number of games as Villanova. Compared with Villanova, this team has more players, then the figure becomes more complicated. But this graph does show pretty high number of assists between players. The number of assists between player Jeffrey Carroll and player Jawn Evans is 110, higher than that of Villanova. And for the players who get the most assists from other players are Brandon Averette and Jawun Evans since they all got assists from other six different players. But we can easily find there is only one connection between player Tyler Underwood and player Will Jerhand. Obviously, Tyler is a new player or bench player in this team.

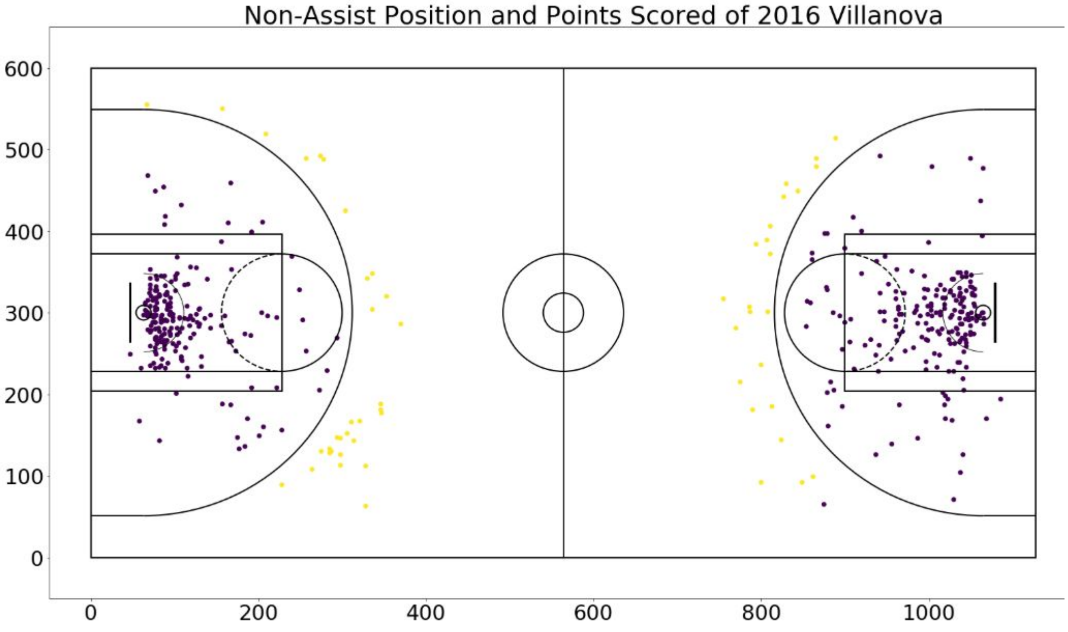


Total Assists between Pairs of Players in Oklahoma

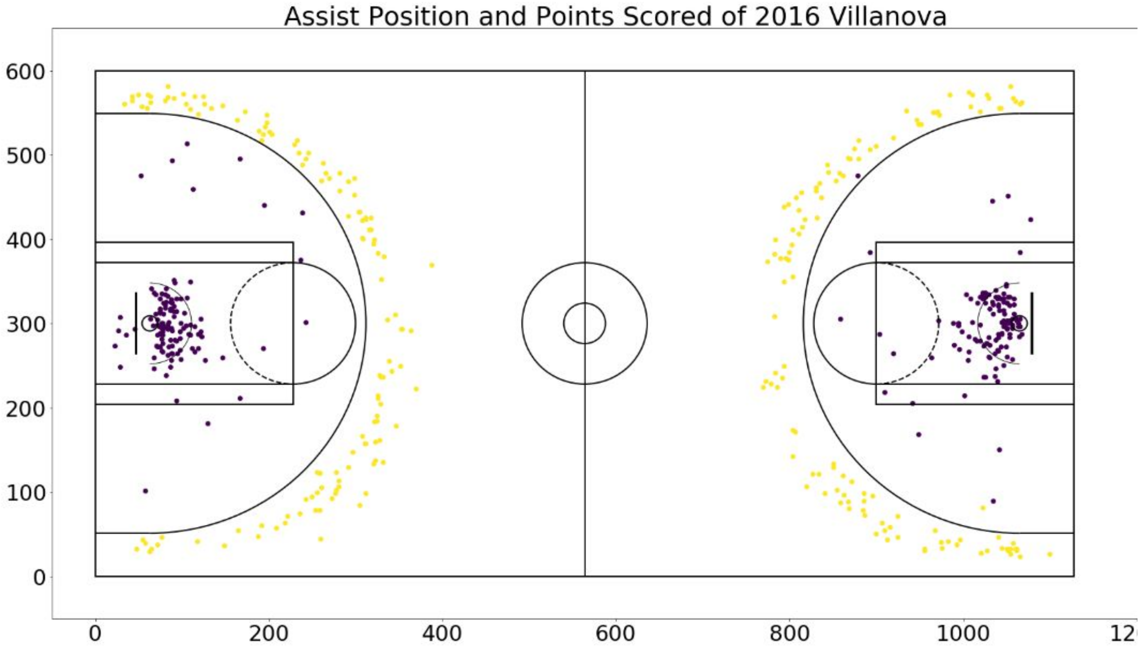
By the above figure, we can still find the connection between Jeffrey Carroll and player Jawun Evans is best in this team that they have 120 assists.

4.2.2. Assists Position

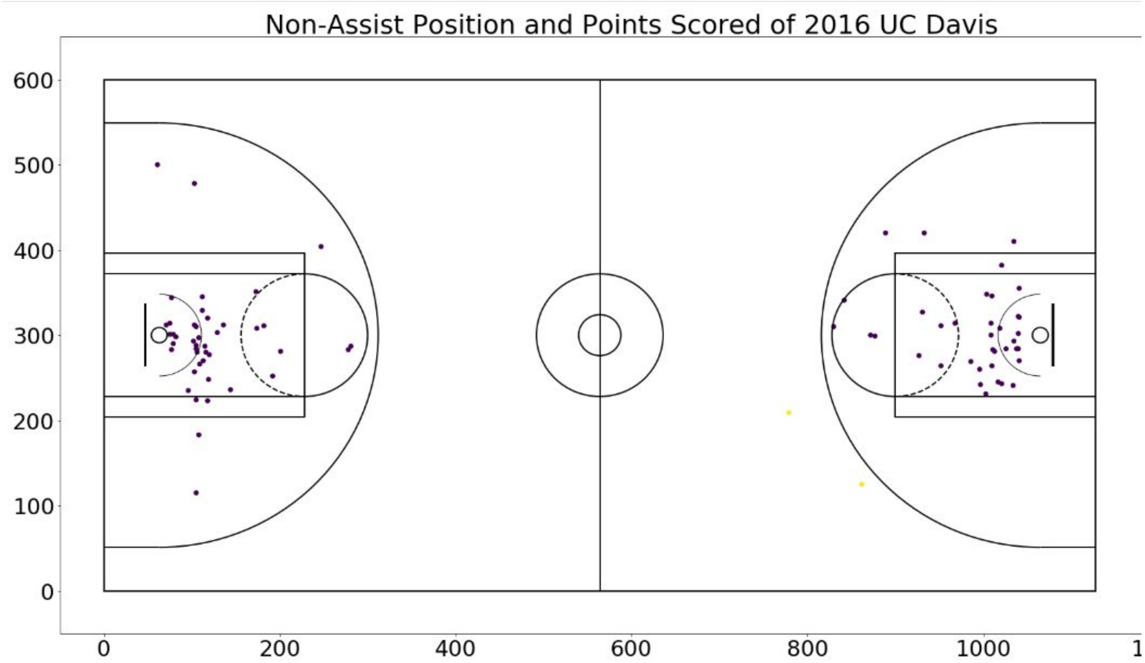
We plot the Villanova and UC Davis teams season 2016 score position for points gained without/with assist to compare the difference. In the following graphs, the yellow dots mean “three points” and purple dots mean “two points”.

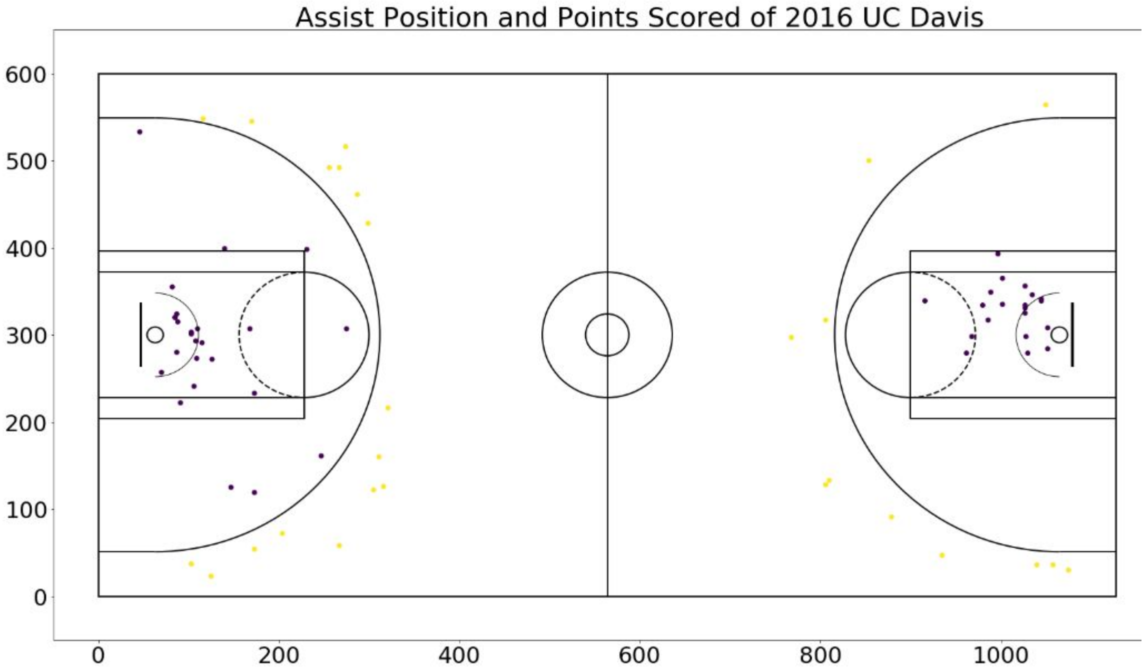


From the above graph “Non-Assist Position and Points Scored of 2016 Villanova”, we can see that there are more purple dots and less yellow dots. This suggests that during games, it’s easier for players to score from a short distance (two points) than a long distance (three points) if no one assists them. This is understandable because there will be more room for the opposite team to block the attack if players shoot the ball from a long distance than a shorter one.



From the above graph “Assist Position and Points Scored of 2016 Villanova”, we can see that there are a lot more yellow dots but fewer purple dots than the previous graph (the one without assist). This demonstrates that three points are much more manageable with organized teammates’ assist. On the other hand, if there isn’t much wiggle room for players to move around (i.e., within the three-point line), scoring will rely more heavily on player’s individual move than teammates' assists.





There are much less dots regardless of color in both UC Davis graphs because NCAA uses elimination matches and UC Davis didn’t play much games in 2016. But even with fewer data, the comparison of “Non-Assist Position and Points Scored of 2016 UC Davis” and “Assist Position and Points Scored of 2016 UC Davis” still follows the two findings we observed from Villanova’s graphs:

(1). A three-point field goal is more manageable with teammate's assist.

(2). A two-point field goal relies more on player’s individual play than teammate’s assist.

5. Obstacles/Future Plan:

5.1. The result of ANOHT is not as good as expected, maybe we need other technique to help us convert numerical variables into categorical variables.

5.2. We try to find connectivity in the directed network (4.2.1). But when we build the contingency table for either Villanova or UC Davis, we get a non-square matrix. We suspect it’s because some player only assists others and never was reciprocated. We will look further into this and find how to calculate the connectivity in 4.2.1. Graphs.

5.3. The computation of mutual conditional entropy: what’s the relationship between joint entropy and marginal entropy, does the mutual conditional entropy introduce by professor always range between 0 and 1, and 1 means they are actually independent?