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 every single basketball game. With this information and analysis, we will then have better understanding what is happening on the court. This understanding is meaningful for the team and coaches since they can adjust their training and strategies accordingly. The information and analysis is also beneficial for the public and players since it can be used to quantify the performance and indicate how to improve. We found really detailed information in our dataset, and, hopefully, our project can serve as a demonstration of the value of the data and analysis.

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:

Analysis of Histogram (ANOHT) provides 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. Our primary interest is that how the position of a player(F, C, G) would affect the number of points he can get. Thus, we applied ANOHT to the points, then color-coded the position of the player. One thing worth noting is that, since the sample size is large, we subset our dataset and firstly focused on the games in season 2017 - 2018 and teams which were the top eight or the last eight.

ANOHT also 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. 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.

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:

4.1.1. ANOHT

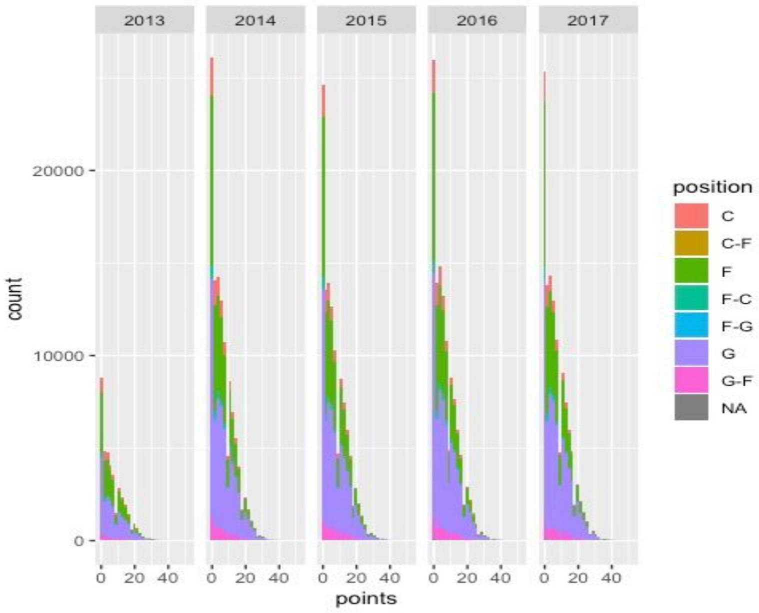


Figure 1. The histogram of points colored by position, seperated by year

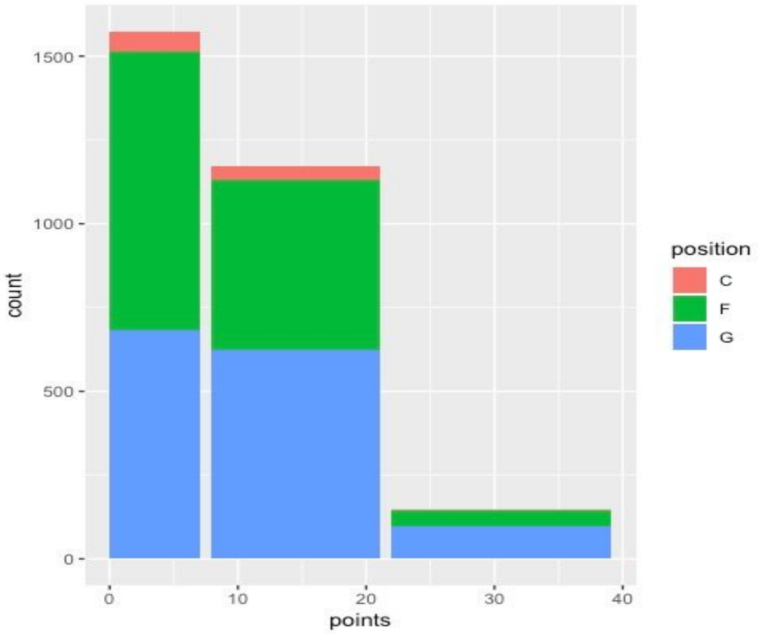


Figure 2. The histogram of points colored by position for the final 8 teams in 2017

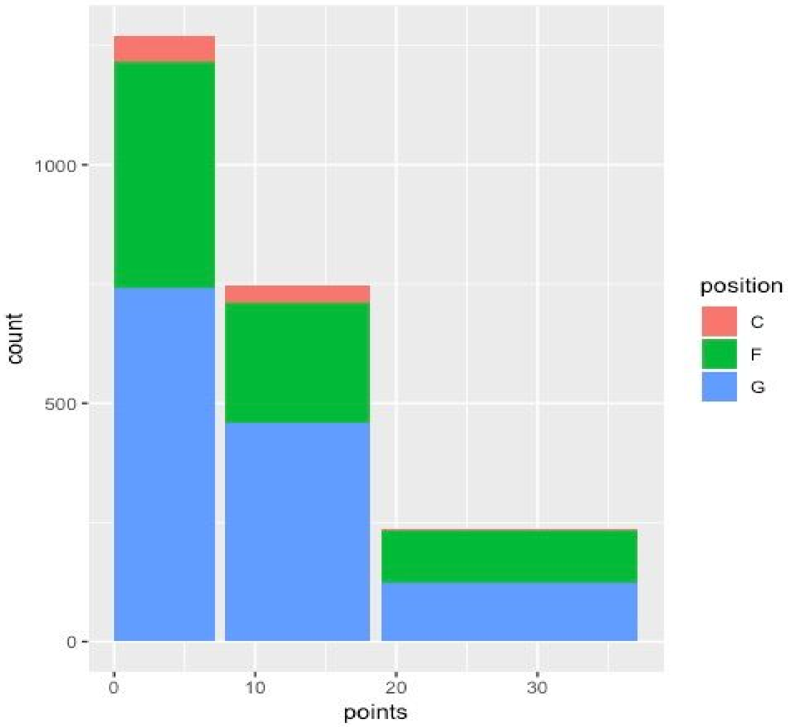


Figure 3. The histogram of points colored by position for the random 8 teams lost in the first round of tournament in 2017

There are 3 groups calculated by histByDess(), which is based on hierarchical clustering. As we can see from figure 2 and figure 3, the major difference between a good team and a bad team is the ratio between guard and forward for scoring high points(The rightmost group on the graph). The guards in a good team are more likely and able to score high points than the forward in the same team. However, the guards in a bad team are equally likely to score high points as the forward in the same team. Furthermore, forwards in a good team are more likely to score lower points than guards, whereas the bad teams show an opposite trend. In conclusion: guards should carry more responsibilities to score more points, and forwards should score less points in a good team.

We also 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.

4.1.2 mutual conditional entropy

Since we have many different methods to try, at this point, we just used the data of the top 8 teams for experiments. In future analysis, we will apply our methods again on the data of the last 8 teams and make comparisons.

We applied ANOHT on many other interested features, such as the player's weight, height, playing time, number of rebounds, number of assists, etc. We then utilized the results from ANOHT and tried to compute the association between different features. The results are shown below in the mutual entropy matrix:

	point	height	weight	rebound	assist	position	hometeam	starter	play_time
point	0								
height	0.9856061	0							
weight	0.9728881	0.6289749	0						
rebound	0.8570826	0.9658106	0.9599889	0					
assist	0.917535	0.9428997	0.9455295	0.9576815	0				
position	0.9817466	0.5298458	0.5043793	0.9705635	0.914525	0			
hometeam	0.9988951	0.9995757	0.9985211	0.9990248	0.9987373	0.9990532	0		
starter	0.8129512	0.9902254	0.9475867	0.8650534	0.8730711	0.9741856	0.9999076	0	
play_time	0.7422297	0.9569335	0.9515628	0.8364982	0.8462652	0.9507457	0.9990948	0.6934125	0

Since the matrix is symmetric, we just need to show half of the matrix for the purpose of cleanliness. Even though we are not completely certain about our findings, we proved that the mutual conditional entropy presented in Dr. Hsieh's paper[citation] is slightly different from the traditional way. And here we are using the formula presented in Dr. Hsieh's paper[citation] to compute the mutual conditional entropy.

We believe that the mutual conditional entropy is a real-valued number from 0 to 1, and lower entropy means higher association. Also, because two features with high association has low mutual entropy, we are able to regard the mutual entropy as the distance between these two features. Thus, the lower mutual entropy means higher association and closer distance between two features. And therefore, we can also regard this mutual entropy matrix as a distance matrix, and generate a graph based on the distance matrix and have further analysis.

There is much information and insights we can have from this matrix. For example, the feature "hometeam" has mutual entropy with all other features close to 1. This implies that the players' performance is hardly affected by the fact of whether they are playing in their own venue. Another example is the feature "point". The mutual entropy of "point" shows that it has relatively high association with "rebound", "assist", "starter", and "play_time". This also makes sense since all these four features can affect the performance of a player and thus the number of points he can get in that game.

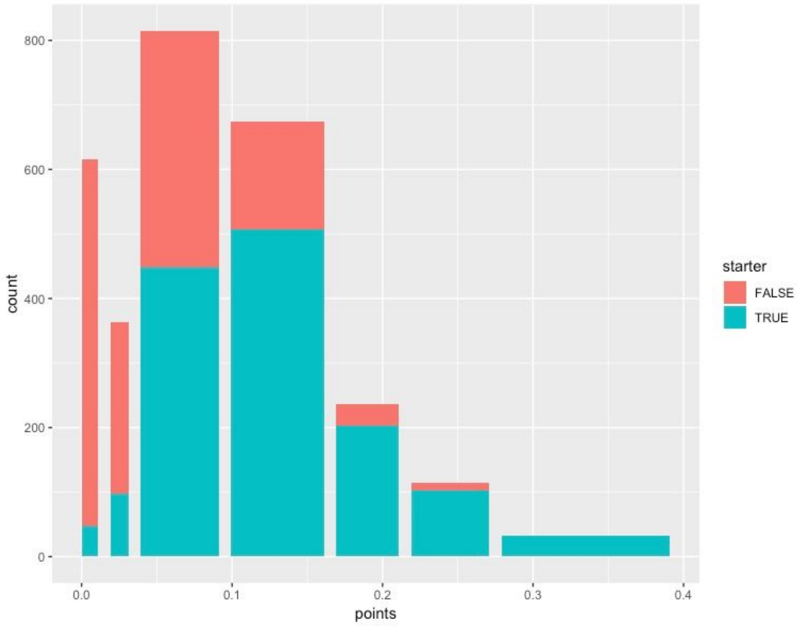


Figure4. Points made by starters and non-starters

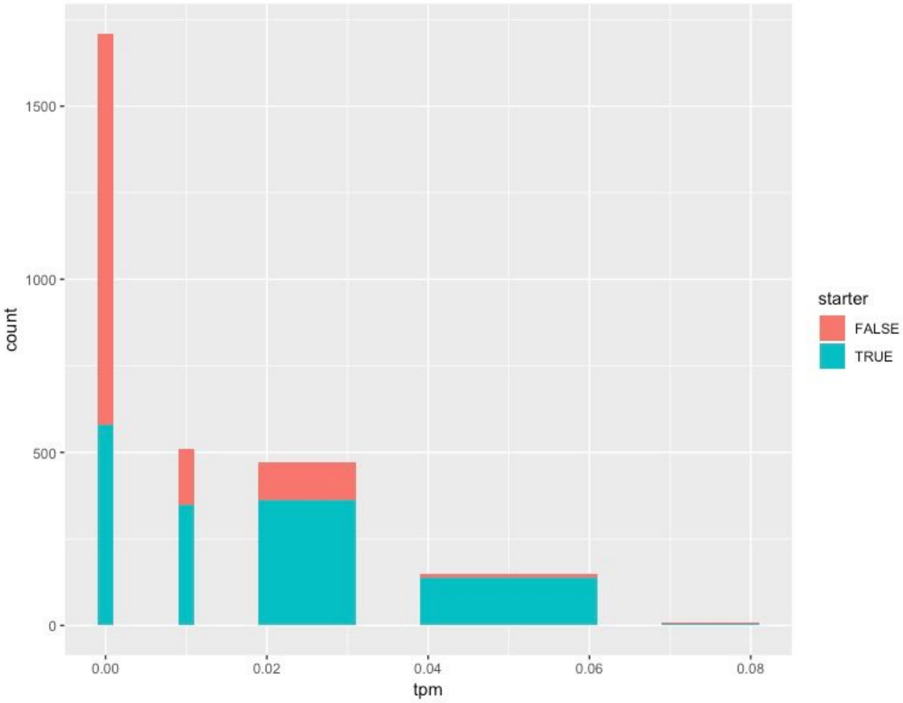


Figure 5. Three points made by starters and non-starters

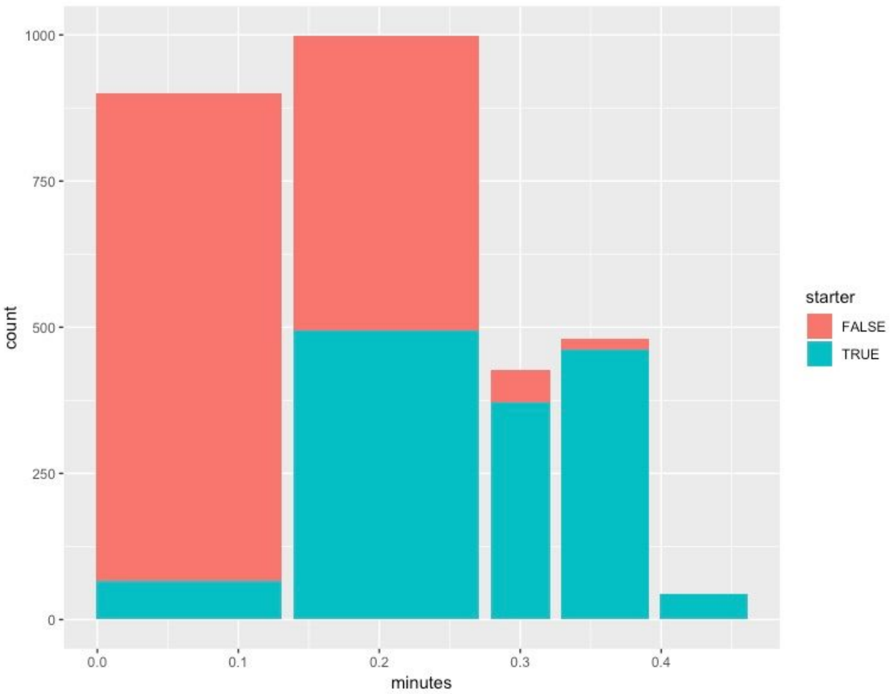


Figure 6. Time played by starters and non-starters

[Term to clarify: TPM(Three points made), starters: player who started the game.]

Figure 4 to 6 shows some categories that have stronger associations with starters. As we can see from the graph, the starters have more time to play the game, score more points, get more rebounds(not shown in this report), and make more three pointers. On the other hand, it reflects that NCAA does not utilize bench players well; they do not get enough time to show themselves. Therefore, having much less contribution to the game. Since we haven't look at the teams that lost in the first round, maybe the bench players from the those teams even have a smaller chance to play the game. Then we can have a contrast, and say "better team use the bench players more and better than weaker team." But in general, bench players are much less efficient than starters.

4.2. Player-Player level:

4.2.1 Player Assists Network

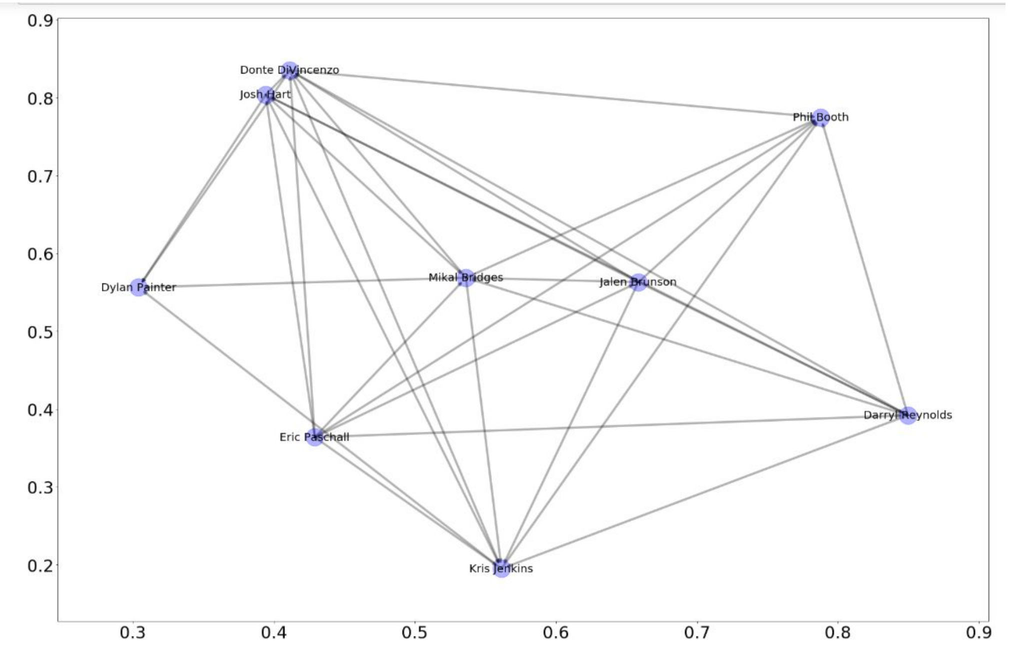


Figure for Villanova

For the figure of Villanova, we can find that each player has at least four links in or out and the entire graph is connected.

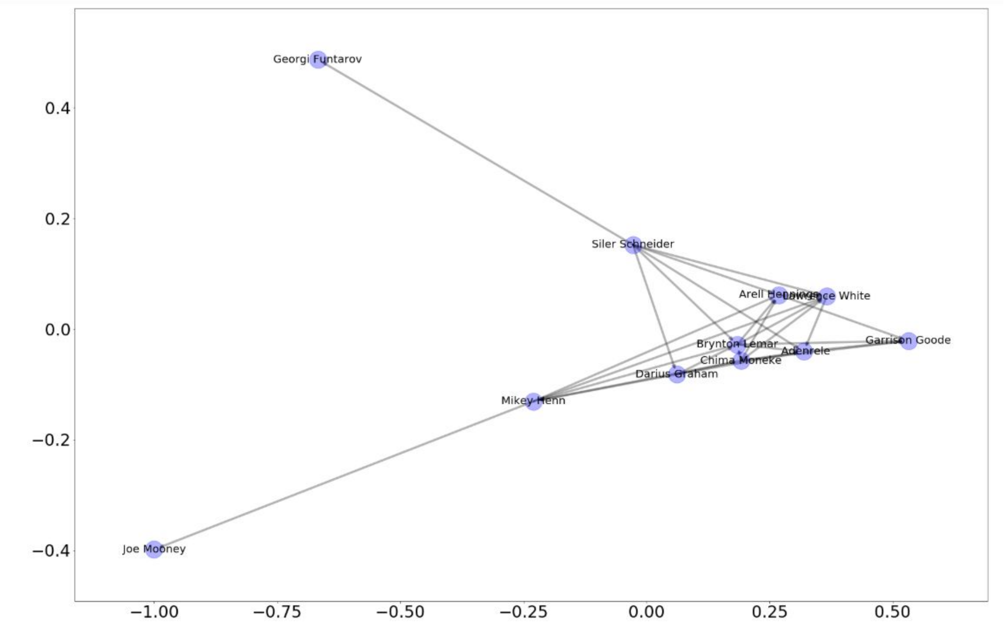


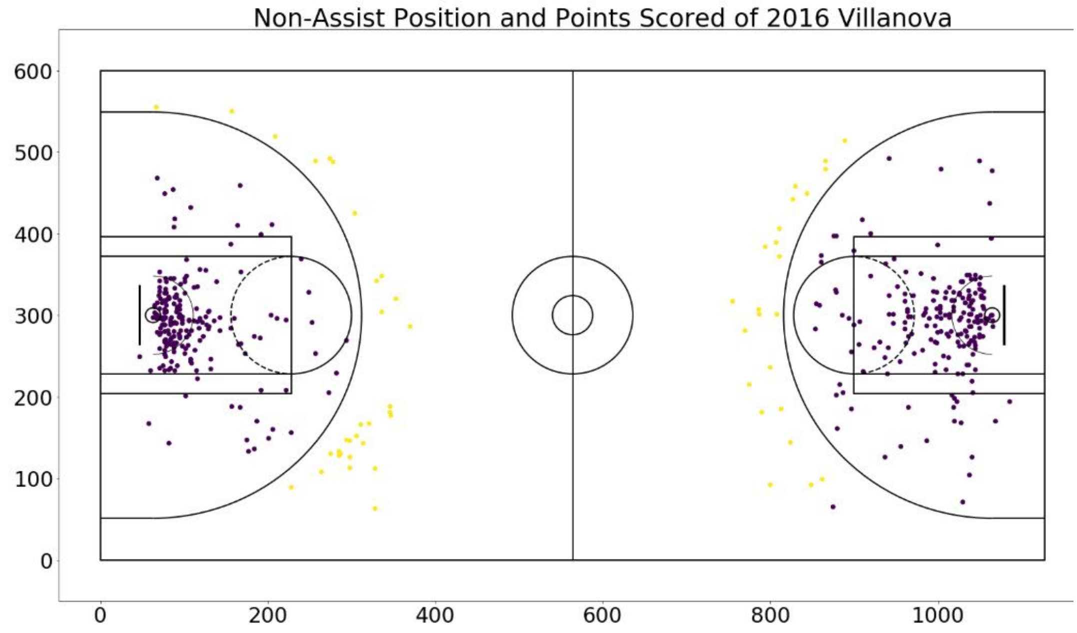
Figure for UC Davis

For the figure of UC Davis, some of players have four or five links with others, however, the rest of them only have one or two edges.

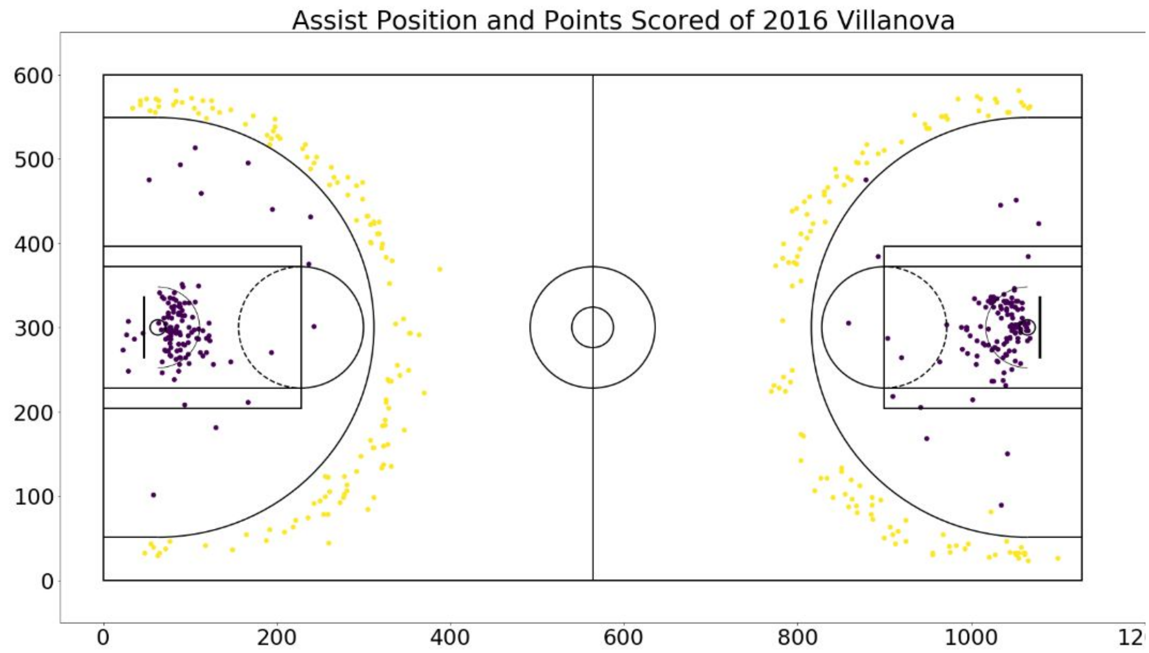
By comparison, the team of Villanova has better teamwork than that of UC Davis. We can also find that although some of player from UC Davis has pretty high personal abilities and skills, they do not work with others. Therefore, the connection between players in a team can potentially be an important factor which affects the winning rate.

4.2.2. Assists Position

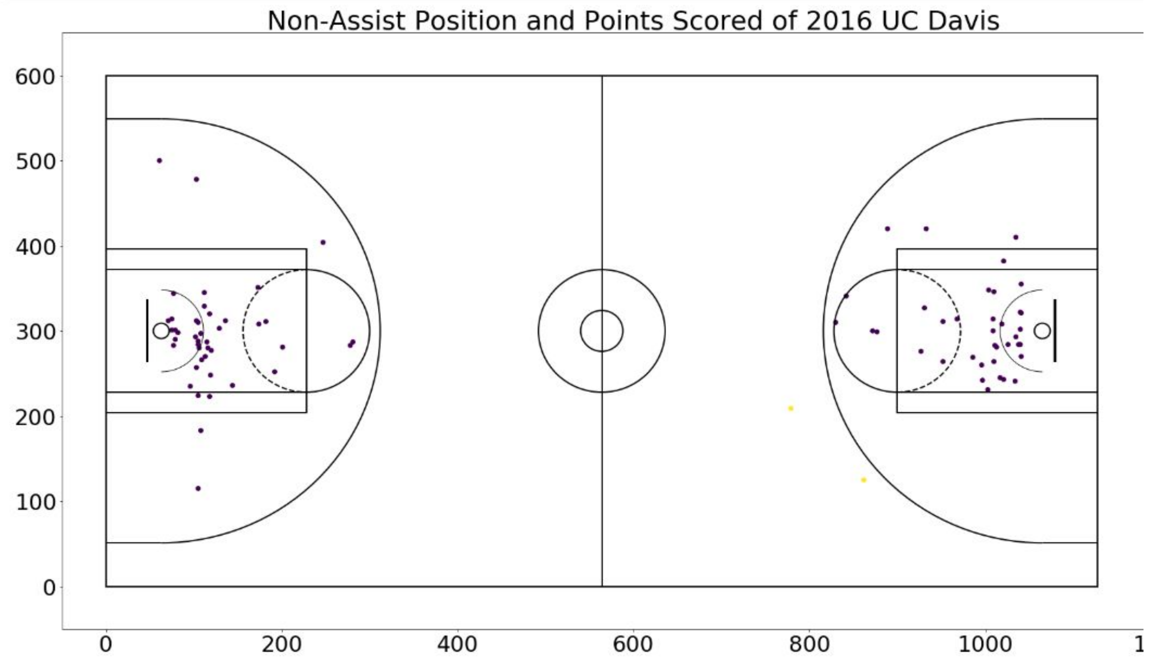
We plot the Villanova and UC Davis teams season 2016 score position for points gained without/with assist to compare the difference. In the following graphs, the yellow dots mean “three points” and purple dots mean “two points”.

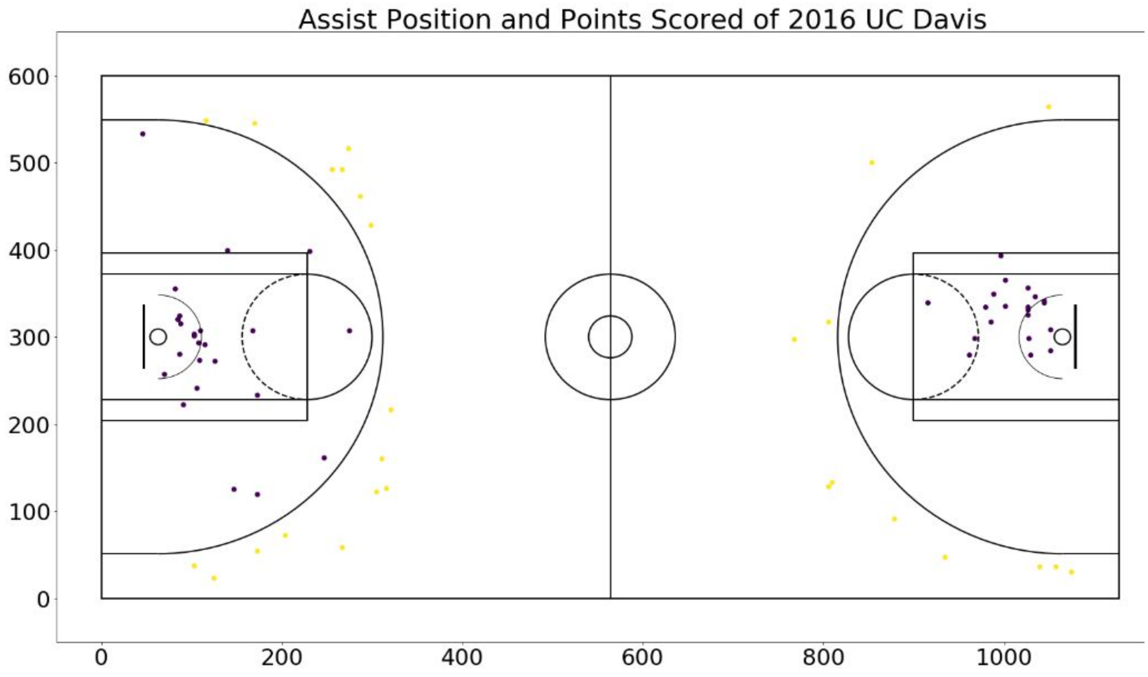


From the above graph “Non-Assist Position and Points Scored of 2016 Villanova”, we can see that there are more purple dots and less yellow dots. This suggests that during games, it’s easier for players to score from a short distance (two points) than a long distance (three points) if no one assists them.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