Data Analysis Project

Analysis of the NCAA March Madness Tournament’s Team Selection

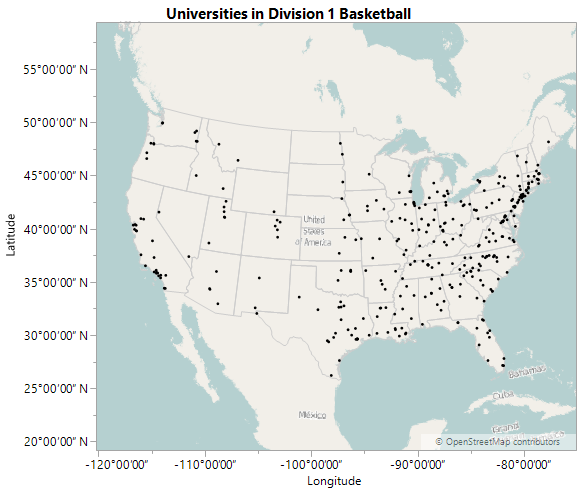
Project by: Jackson Ridge and Alex Cook

STP420

Neil Hatfield

Data serves as an important tool when analyzing the performance of teams in collegiate sports, and the most important application is in the ordering of teams from best to worst. Different analysts utilize different qualities of a team to determine the team’s ranking. Some people emphasize overall record, while others emphasize performance measures such as overall rebounds or total number of turnovers per game. A team’s performance, however, is not the only measures that analysts use to rank teams, and many opinions are biased. Teams that have won the national championship in the past are considered to have an extra advantage, in the playoff experience of the players or coaching staff and can be ranked higher than their performance suggests, while lesser known teams can be underestimated due to lack of experience.

There are four questions determined to be the most applicable to the data set and most relevant to college basketball fans.  The main question is, did the selection committee accomplish their goal of picking the most capable basketball teams?  Expanding on this topic, is there a bias of the committee on selecting teams between major and mid-major conferences?  These questions are focused on whether the teams that made the tournament have significantly higher statistics in the most relevant categories (such as winning percentage) over teams that did not make the tournament.  This led to another interesting question, of if there are basketball statistics that show the least difference between tournament and non-tournament teams.  Finally, the last question that will be analyzed is which individual statistic ranks the teams most closely to the rankings devised by Ken Pomeroy, a highly-respected college basketball analyst.

 The ranking systems rank the teams from one to 351 (There were 351 teams in Division One basketball during the 2015-2016 season) with the first ranked team representing the overall best team, and the 351st ranked team representing the worst. The rankings published by Ken Pomeroy will be used, as he is one of the most revered and well known analysts of college basketball. The most current season that was finished, with all the statistics recorded, at the time of beginning this project, was the 2015-2016 season, and these will be the data used in all analysis. The 2015-2016 season will reflect upon the committee’s most current selection method for the NCAA tournament.

Each of these 351 teams (shown on the map above, Hawaii left out) belong to a conference. Conferences are a collection of teams that agree to play against one another. Conferences are formed for financial reasons but also take into consideration geographical location. There are a total of 32 conferences in Division One college basketball. The conferences are further split up into an unwritten divide, these are the “power” conferences and the “mid-major” conferences. Of the 32 conferences, 6 are power conferences, and the remaining 26 are the mid-major. The NCAA does not actually classify conferences in this manner, but spectators, sports broadcasters, and analysts know this stratification exists.

Mid-major conferences are the conferences that are comprised of teams that typically lose more games than they win. These teams are also not historically prestigious basketball programs, although there are a few programs that are an exception to this. Power conferences contain teams that have played well consistently throughout the decades and more often than not have a perennial winning record. Examples of teams that fit this category are Kentucky, Kansas, Villanova, North Carolina, Duke, and various other colleges.

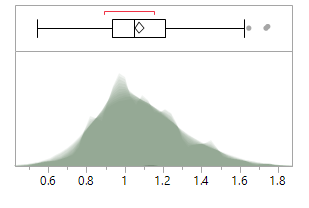
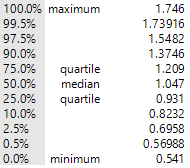
When looking at a school's schedule, they are always comprised of non-conference games and conference games. Most of the games a team plays during the season are conference games while non-conference makes up a smaller percentage of a school's schedule. One reason this is important is because non-conference games heavily impact a team’s strength of schedule, since a team can schedule their non-conference games against any university they choose. In addition, non-conference games give the best indications of how well conferences are comparatively, in order to help determine rankings.

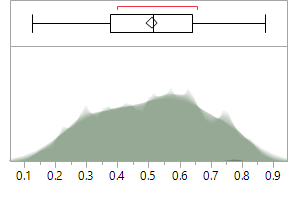
Another facet to look at is teams who made the NCAA tournament via an “at-large-bid” versus the teams that are “on-the-bubble”. The NCAA March Madness Tournament has a total of 68 teams which get into the tournament in two different ways. The first method that a team can get into the tournament through is called an “automatic-bid,” which is when a team wins their conference tournament at the end of the season. The other way a team can get into the tournament is through an “at-large-bid”. An “at-large-bid” is when the NCAA basketball selection committee picks teams based on their performance during that season. These teams are considered some of the best and most consistent teams in all of the NCAA for that particular season. And lastly, teams that are “on-the-bubble” are teams that are on the verge of getting an “at-large-bid”.

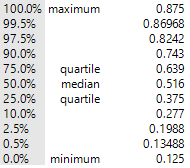
The statistics for each team should play a vital part of how they are ranked, but as previously stated, that is not always the case. “Blue-bloods,” which are schools with a long tradition of having a winning basketball team, are normally favored, although there may be teams that are equally as good, but lack a rich history of basketball or playoff experience.

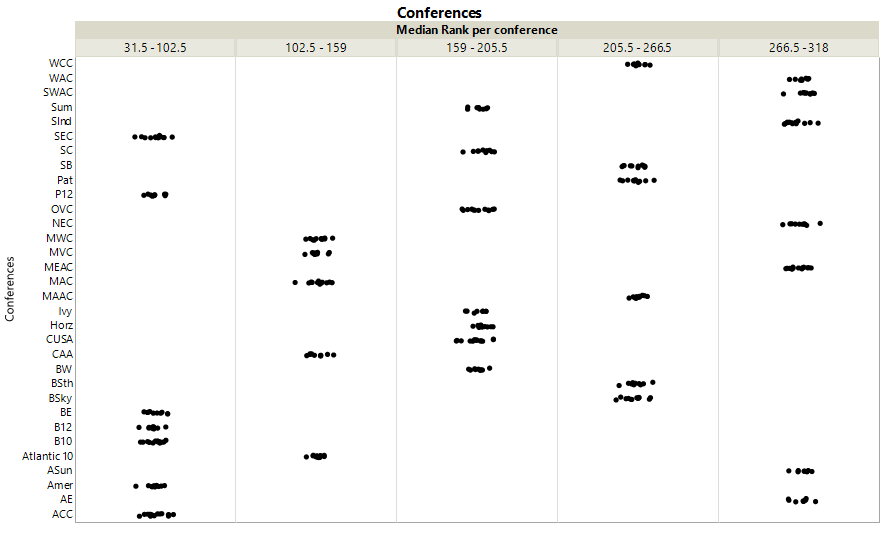
Now that it's been explained how teams are classified, ranked, and awarded in the postseason, it's time to look how each team performance according to the different metrics provided.

The first statistic chosen to be analyzed is the assist to turnover ratio. This is represented as a fraction that is calculated by dividing the number of assists a team has (assists are when someone passes the ball to a player then they score) by the number turnovers a team has (a turnover is when a team has possession of the ball then lose the ball without scoring a basket). This statistic highlights a team’s ball handling skills, and a larger value is more indicative of a better team.

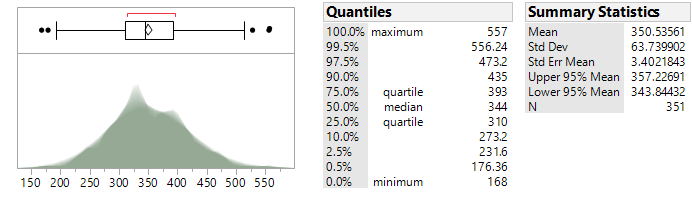
The graphic below displays the distribution of all 351-team’s assist to turnover ratios. There are three outliers present in the data, and belong to Saint Mary’s, Michigan State, and North Carolina, all of which ended in the top 50 ranked teams. Teams that are above the 90th percentile were analyzed (1.3746), because those select few teams are exceptionally good in this category. There are 11 teams who are in the top 50 who meet these criteria, seven teams ranked 51-100, and seven teams who are ranked 101-351. The shadowgram shows that the peak of the turnover to assist ratio is at about one, meaning a large group of teams have just as many assists as they do turnovers.

A team’s win-loss percentage is the most obvious indicator of how decent a team might be, and this is the number to games a team wins divided by the total number of games a team plays (anything lower than .500 means the team lost more games than they won). The 90th percentile or higher is a good place to start when looking at this metric. The 90th percentile or higher is contained in the interval [.743,1]. In this interval, there were 23 teams in the top 50, nine teams ranked 51-100 and 5 teams who are ranked 101-200. When looking at the shadowgram there are no noticeable prominent modal clumps that would indicate there is some other factor to take into consideration.

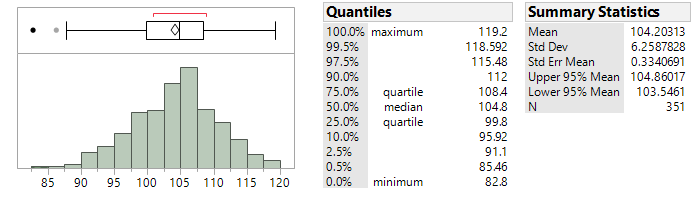


A combination of both assist to turnover ratio and winning percentage was considered to see which teams fit both conditions. There were two noticeable exceptions that had a high win-loss percentage and a high assist to turnover ratio and did not make the tournament, Saint Mary’s(CA) and the University of Alabama-Birmingham (UAB).

All 351 teams can be stratified into conferences. Each conference differs in size, the sizes of each conference all fall within this interval [8, 15]. When examining a conference, the best measure of strength is to consider the individual universities. Having many highly ranked teams in one conference is a strong indicator of the overall conference strength. One way to determine this is to look at the ranking of each team and find the median. The median is a more useful statistic to consider than the sample arithmetic mean, since the rankings of each In the graphic above, it is apparent that the conferences with the highest median ranking are the PAC-12, the ACC, the Big 12, the Big 10, the SEC, the American Conference, and the Big East. This is because the farthest left column contains the conferences with the lowest median ranking. These all happen to be the power conferences, except the American conference, which is considered mid-major. Thus, teams in the power conferences show tendencies to be better ranked than teams in other conferences.

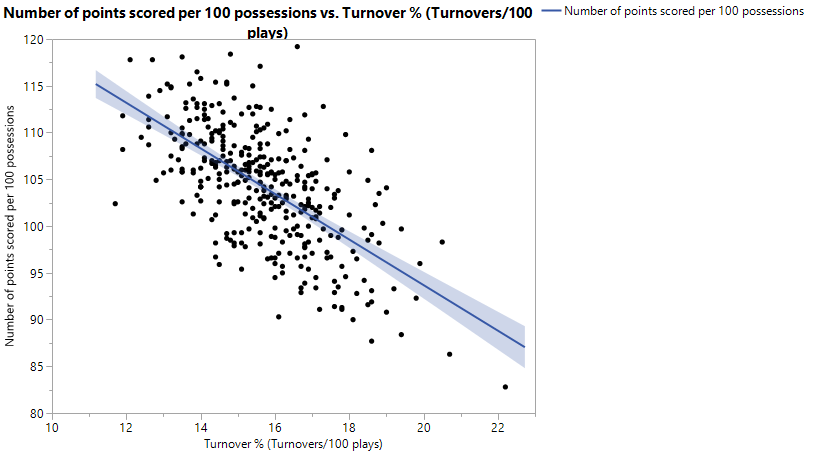
 Another important offensive metric is the number of offensive rebounds a team gets. An offensive rebound is when a team regains possession of the ball back after a missed shot, while playing offense. Offensive rebounds give a team another chance to score, so the higher the number of offensive rebounds in a season, the more likely a team is to have a greater value of points scored in their games. When looking at the box plot for the number of offensive rebounds in a season, three teams stood out: North Carolina, West Virginia, and Morehead State. These teams were considered outliers on the right side of the box plot meaning their values for offensive rebounds were considered outliers for being very high.

After looking at just the distribution, the top 90th percentile of teams were considered to see which of those had a winning record over .563, which is the lowest winning percentage a team has ever had during the regular season while still making the tournament with an at large bid. That leaves the following teams which meet the qualifications and also didn't make the tournament: South Carolina, Georgia Tech, Florida, George Washington, SDSU, Nevada, VCU and BYU.

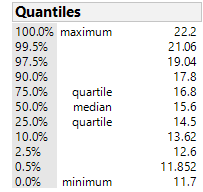
Probably the most useful metric out there on determining how good a team is at scoring is the number of points scored per 100 possessions. This metric adjusts the scoring of a team based on number of possessions, since it takes into consideration the pace of the game and how fast teams might score points. It looks at the first 100 times each team has the ball and sees how many points they scored, this makes it easy to compare each team due to the variation in how fast a team plays. While looking at this metric, the sample arithmetic mean (SAM) for the entire NCAA was found and compared to each of the conference’s SAMs, to see which percentile to further see which conference is the strongest for this metric.

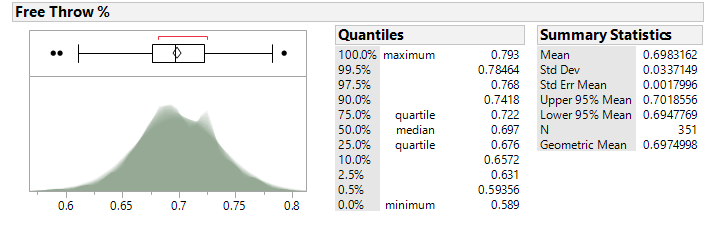
The SAM for this metric is 104.20313 points per 100 possessions per team. This is right below the sample median which happens to be 104.8. When looking at the individual SAMs for each conference only 11 of the conference's SAMs fall below the SAM of all the entire NCAA. Those 11 conferences were filtered out since having a lower number for this metric means that the teams are not productive when they have the ball. Again, the top 90th percentile of all the teams (so teams that were 112 or higher) was considered and teams that were in those 11 conferences that were below the overall SAM were also removed. The universities that were in the top 10th percentile and were in the remaining 21 conferences were: Saint Mary’s, BYU, North Florida, Belmont, Winthrop, High Point, and Eastern Washington.

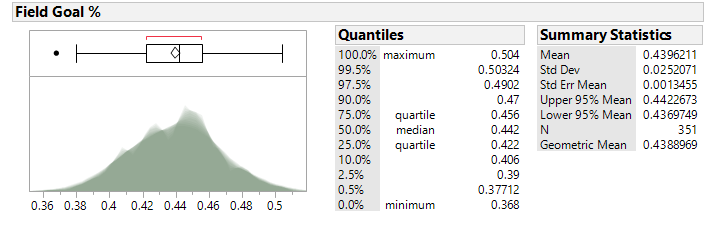
From the metrics examined so far Saint Mary’s and BYU (in the West Coast Conference) have both shown up more than once meaning these teams performed well.



Next, both percent of turnovers per 100 possessions and number of points per 100 possessions were considered. Just like number of points per 100 possessions, the percent of turnovers per possession looks at how many times a team turns over the ball for every 100 times they have the ball on offense. This is then converted to a percent, so if a team has a value of 21, that means a team turns over the ball 21% of the time for per every 100 possessions.

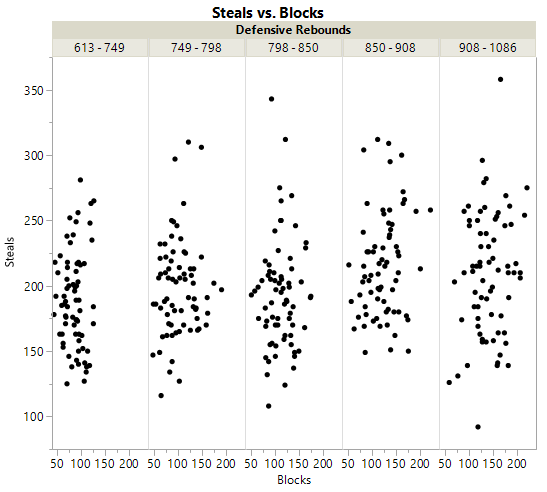
In the graph above, there is a clear correlation between having a low number of turnovers and a high number of points scored. Since this is the case, are going to look at median for both are and then find the teams that fit both criteria. The quantiles for the % of turnovers is below, while the number of points per 100 possessions is already in the paper and was shown earlier. Since its turnovers, a smaller number is better so the 10th percentile or lower and the 90th percentile or higher for the other metric will be considered. According to the data, the teams that meet the criteria for both and didn’t make the tournament are: Saint Mary’s, Houston, and Oakland.

Since the outcome of a basketball game is ultimately determined by score, the number of times a team makes or misses a basket determines whether the team wins a game, therefore field goal percentage and free throw percentage are both important metrics to consider. The following visualizations show the distributions for each of these metrics, and the 90th percentiles of each are the most important regions.



For both the free throw percentage and field goal percentage the Sample Geometric Mean was considered, because each team’s percentage was calculated by the total number of baskets made divided by the total number of baskets missed, and each team had different values for their numerator and denominator. By looking at the SGM for both the teams that are at the geometric mean or above for both will be filtered out, leaving teams that fit both qualifications. Then, winning percentages of .743 and higher were considered, which is the composition of the 90th percentile. The teams that fit all these criteria and did not make the tournament are: Princeton, UAB and Valparaiso.

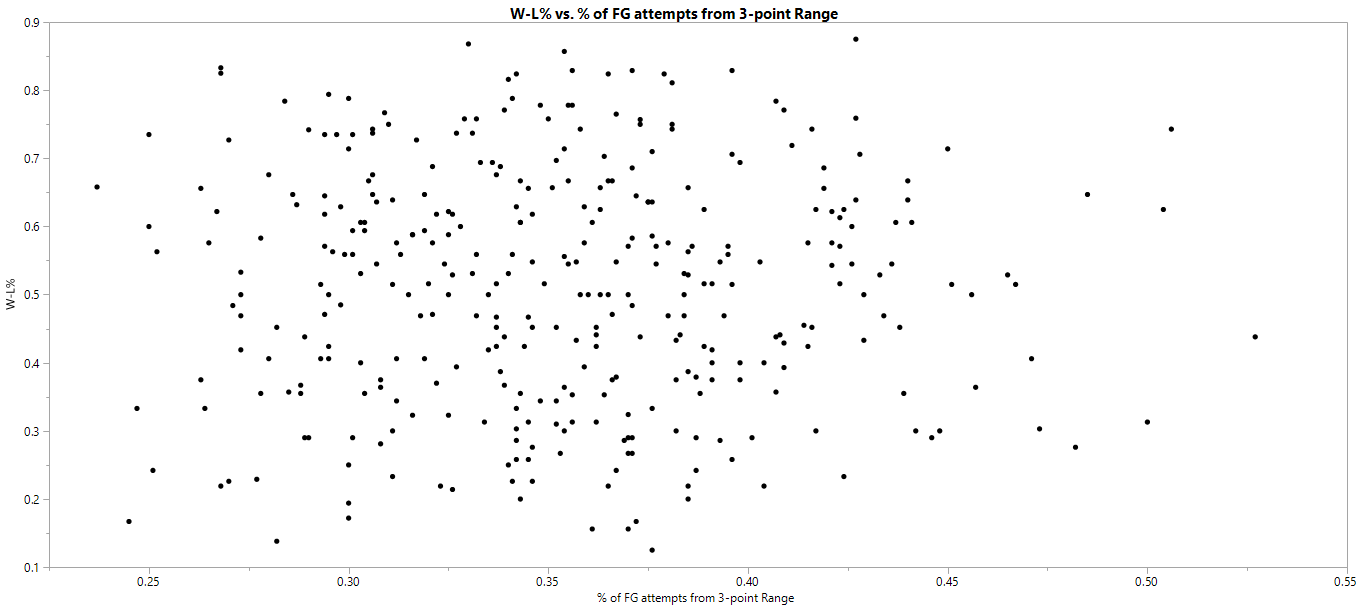
Scoring is important, but a solid defense also helps a team win games. Steals, blocks and defensive rebounds are all metrics indicate how good a team is defensively. A block is when a player prevents the opposing team from scoring a basket by swatting the ball out of the air and preventing the other team from scoring. A defensive rebound is the same exact thing as an offensive rebound except a team is getting the ball after the opposing team missed a basket.

The graph below shows those three variables. The visual display is split up into five sections according to how many blocks each team had total in the season. The far left being the teams with the least number of blocks while the far right shows the teams with the largest number of defensive rebounds in a season. In each section the dots that are farthest to the right in their section and closest to the top are the teams that performed well in the steals and defensive rebounds category. The far-right section shows teams with the highest numbers of defensive rebounds, and from this section, teams that are above the SAM for both blocks and steals were considered, with the SAM for blocks and steals are 113.29345 and 201.05128 respectively. When looking at the teams who didn't make the tournament and are above both SAMs and are in the far-right section of the graph and have a record of at least .563, they were: USC, Florida, Nevada, SDSU, East Tennessee State, Old Dominion, BYU, Northern Illinois, George Washington, Monmouth, Oakland, Valparaiso, Vermont, and UC Irvine. These are all teams that showed to have a relatively strong defense in comparison to every other team in division 1 of the NCAA but did not make the tournament.

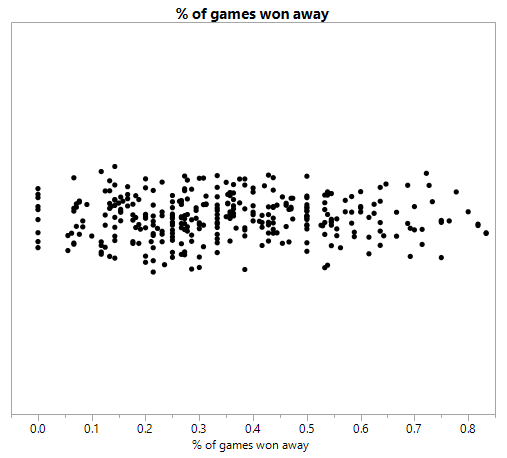
We’ve already looked at turnovers when looking at assist to turnover ratio, but when looking at both turnovers and personal fouls one can determine how disordered and inconsistent a team is. When having a bigger number for both category, this is very detrimental to team performance. The SAMs for both were considered to determine which conferences have both of their SAMs above the SAMs of all the teams. This will help determine which conference has the least sloppy teams. The SAM for turnovers is 412.11681 and the SAM for personal fouls is 629.16809. The SAM of each conference’s win-lose percentage was also taken into consideration, and conferences that are at or over a .500 win-loss percentage were chosen. The conferences that fit these qualifications are: The Big 10, the ACC, the AAC, the Atlantic-10, the Ivy League, the Summit League, and the Colonial Athletic Association.

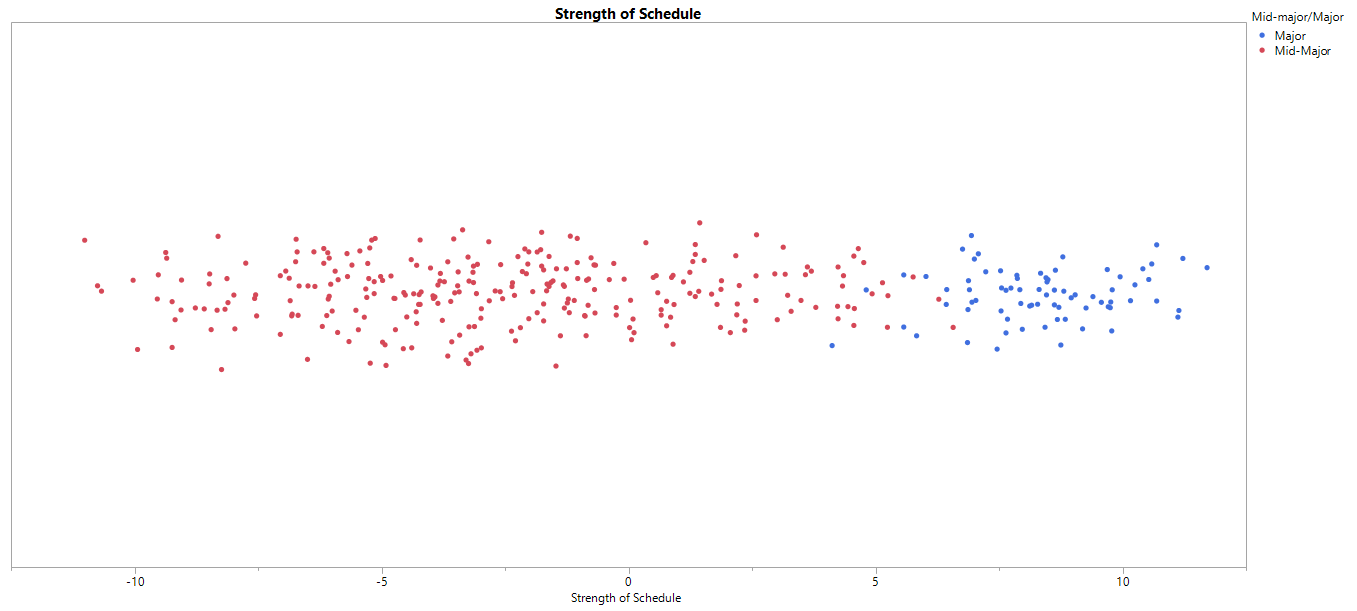
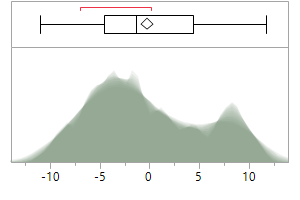
From how bad the each conference is at ball handling and a conference’s median rank, the conferences that should be considered are the Big 10, the ACC and the AAC. This would indicate that these three conferences have the highest-ranking teams and play the most refined game of basketball.

One metric that most people assume is beneficial to a team is whether or not they decide to shoot from the three point range, or shoot an ordinary basket worth two points. It is widely thought among basketball fans that the more three point shots a team attempts, the more likely they are to win. The graphic below shows on the x-axis shows how many of their shots attempted were from three point range, while the y-axis shows a school’s win-loss percentage. The points do not visually seem to follow a trend and appear to be variously scattered across the visual display. From this it can be concluded that there is no clear relationship between the percentage of 3-point field goal attempts and the win loss percentage of a team.

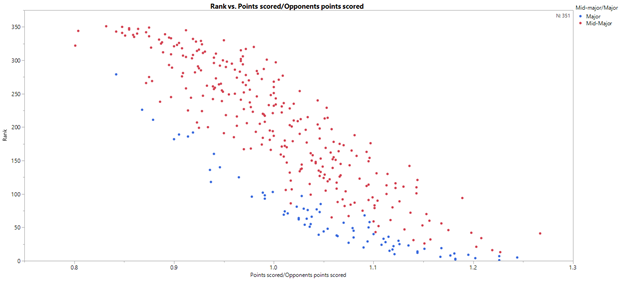


A team’s percentage of wins for an entire season has already been considered, but breaking down those wins even further is important beause each win has a different weight to them. Winning an away game is deemed more difficult than a home game due to home court advantage. Even the officiating and the crowd can have an effect on an away team’s performance. “Following episodes of spectator protests (because the referee blew his whistle and it was bad for the home team) there was a consistent pattern of improvement in home team performance accompanied by a decline in visiting team effectiveness” (Greer 25). Out of the top 50 teams who has the best winning percentage for away games, 25 of them made it to the NCAA tournament. And of those 50 only nine of those teams were from a power conference.



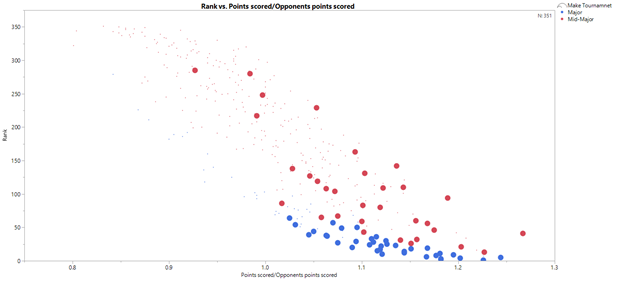
 Strength of schedule is a really important factor when determining how good a conference is and whether a team is going to make the NCAA tournament. Because of this, a team can 20 out of their 28 games, but with a bad strength of schedule, the wins do not have as much relative value as wins for a team with a high strength of schedule. This figure below shows the shadowgram of the distribution of strength of schedules of all teams, and two distinct modal clumps are visible, one centered around -4 and one centered at +8. The most likely cause for these clumps would be the separation by major and mid-major conferences. To investigate further, a visualization comparing the strength of schedule to conference type was created, which further supports this conclusion. The majority of teams with stronger strength of schedules, located in the right side of the chart, are predominantly from major conferences.

When looking at a team’s strength of schedule it was found that there was a divide between teams that are in a power conference and teams that are in a mid-major conference. In the graphic being further to the right means the team has a strong strength of schedule, while being further left means the opposite. Almost every team in a power conference (blue) has a more difficult schedule than those teams in a mid-major conference (red). These weak schedules could be a contributing factor to why teams who have high win-loss percentages and are on the bubble do not always make it into the NCAA tournament.

One of the more promising metric that helps to determine the quality of a basketball team is the ratio of points scored to points opponents scored, as it combines both the offensive and defensive capabilities of a team. The data visualization below compares this ratio to the rank of each team, with the color of each point representing each team’s conference as major or mid-major. The trend in the data is to be expected, with teams with higher points scored to points scored by opponent ration having the lowest and best ranks. By the coloration, there is a clear pattern of major conference teams being higher ranked than mid-major teams.

The data visualization below is the same as the previous, except with the larger marks represent the teams that participated in the NCAA tournament. All but four of the teams in the tournament had points scored to points scored by opponent’s ratios greater than one. With one final modification, emphasizing the teams that made the tournament. There almost seems to be an equal number of mid-majors and power conference schools. To be exact there were 32 mid-major schools that made the tournament and schools from a major conference that made it in. This might appear that both are equally represented, but when examined more closely, there appears to be a significant effect made by the committee on these numbers.

As mentioned previously, there are a total of 32 conference, five major, 27 mid-major, which means that 27 mid-major schools automatically make it into the tournament and 5 schools from a power conference automatically make it in. Out of the 32 mid-major schools that made it into the tournament only five made it in via an at large bid. Those five teams only came from two conferences, the Atlantic 10 and the AAC. The remaining 36 teams were all from power conferences, but out of this 36, 31 of the teams made it into the tournament by receiving an at-large-bid.



Overall there we would say that there is a bias that leans towards strength of schedule. The teams with the strongest strength of schedule also happen to be the teams who are in power conferences. However, there are some mid-major teams who repeatedly show up in some of the metrics that we chose, but those were not enough to override strength of schedule. Based on this we can say the committee did as best of a job as they could base on the data. There were teams that did have semi-high stats in particular areas, but they was not significant enough to classify them as a team that should have made the tournament. The committee might have shown some bias during the 2015-2016 season, but ultimately the teams that were left out did not have good enough metrics to get an at-large-bid for the tournament.

Citations and Sources

“2015-2016 College Basketball Stats.” Sports Reference College Basketball. Web. http://www.sports-reference.com/cbb/seasons/2016-school-stats.html

“2016 Pomeroy College Basketball Ratings.” Kenpom. Web. http://kenpom.com/index.php?y=2016

Greer, Donald L. “Spectator Booing and the Home Advantage: A Study of Social Influence in the

Basketball Arena.” *Social Psychology Quarterly*, vol. 46, no. 3, 1983, pp. 252–261., www.jstor.org/stable/3033796.