**March Madness Report**

March Madness is the nickname for the NCCA Division I men and women basketball tournament played at the end of every basketball season. It has the nickname March Madness because there’s always a lot of hype, excitement, epic victories, crushing defeats and it’s very entertaining as it’s one of the most anticipated basketball events of every year. Only 68 teams make the tournament and each of these teams battle it in a win and advance style tournament until a champion is crowned. Fans get in on the “madness” by filling out brackets in an attempt to pick a perfect bracket, or at least better than their friends. Now that sports betting is becoming legal in the United States, a lot more is going to be at stake when watching tournament games than just picking a good bracket which means correctly picking who wins and loses each round. The better the bracket, the more money that could be won. I wanted to do this project to analyze what characteristics of teams matter when filling out a bracket. I wanted to find out which conferences send the most teams to the tournament and if that plays a role in how deep they go. I also wanted to see if the teams with the best stats go the farthest, or if being a 1 seed matters. The questions I seek to answer include: What Conferences sends the most teams to the tournament each year? Do 1 seeds have an advantage to win Tournament? What stats are most important to winning? What are stats like of teams that win the tournament? What’s a profile of the winning teams? Let’s see what I find.

The dataset used to complete this project is from Kaggle called College Basketball Dataset. The data has 5 csv files. Four of the files have 23 columns and they each contain data from Men’s Division 1 college basketball seasons from the 2015 through the 2019 seasons. The fifth csv file is a master file which has all four files combined. Field Descriptions are below:

A screenshot of a social media post

Description automatically generated

A screenshot of a social media post

Description automatically generated

To begin my program, I uploaded the libraries that I would be using to complete the project. Next, I changed my working directory so I could easily upload the data files. Then, I made my notes or a plan of action to refer to as a tried to answer each question. I did this before I did any coding for each question. To answer my first question: What conferences sends the most teams to the tournament, I uploaded all the college basketball data from 2015 through the 2019 seasons in a pandas dataframe. I checked the 2015 dataframe to make sure all the data was loaded properly. I didn’t need to make any changes to the data for my first question. I also didn’t need to do anything with NA’s in my data because I used them to help me figure out what to exclude when I needed to manipulate my data. Then I print out how many teams are in each conference and how many teams qualify for the tournament. I then make a new dataframe that filters the original to have only teams that qualify for the tournament. Next, I print out a bar chart that visualizes the results. The charts for each of the years are in the program. I ran all my code in Python using a jupyter notebook.

For my second question: Do 1 seeds have an advantage to win Tournament, I uploaded the master csv file into the notebook. Next, print out a few rows of the dataframe to make sure the data is all there. After that, I make a new dataframe that filters the master file to contain only the teams that won the tournament each year and I print that out to display the results. For my last questions: What stats are most important to winning, What are stats like of teams that won the tournament, I uploaded the master file again. For this question, I wanted to run a correlation analysis. I deleted columns I felt where unnecessary and added a winning percentage column. Then I ran the correlation analysis and printed out the results. Next, I made a new dataframe with the teams that won the tournament but only included columns based off of the results from the analysis to view the stats. I also printed out the averages of the columns that were important to get a profile of a team that could win the tournament based off the statistics.

What conferences sends the most teams to the tournament? From my analysis, I learned that there are a handful of conferences that send a lot of teams to the tournament. However, the ACC, SEC, Pac 12, Big East, Big 10, Big 12, and the American Conference sends the most teams to the tournament. In the 2017 and 2018 tournaments, the ACC sent nine teams from their conference to the tournament. That’s the highest total number of teams from any of the top conferences. That means that these conferences really make up a big bulk of the teams in the tournament every year. The rest of the conferences only sent an average of one team to the tournament per year. The better basketball programs in the country play in the ACC, SEC, Pac 12, Big East, Big 10, Big 12, and the American Conference. When picking a bracket, it’s probably best to pick teams from these conferences to go farther in the tournament because these conferences send the most teams to the tournament. My project does not analyze how teams get into the tournament. I didn’t include that because outside of the teams who win their conference, the rest of the teams that make it are selected by a committee. There’s no set formula for the way they select those teams so I didn’t go into that. The way teams are seeded is also chosen by the committee. How teams get that is also outside the scope of my project.

Do 1 seeds have an advantage to win Tournament? I found that 1 seeds definitely have an advantage to win it. A number 1 seed won the tournament four of the five years of my dataset. The only exception was Villanova who was a 2 seed in 2016. Villanova won again in 2018 as the 1 seed. Also three out of the five winners came from the ACC; Duke in 2015, North Carolina in 2017, and Virginia in 2019. The other two winners came from the Big East; Villanova won in 2016 and 2018. Earlier I mentioned that these conferences sent a lot of teams to the tournament. The ACC and Big East clearly dominate the winners circle. Another reason why it’s smart to pick 1 seeds to go deep into the tournament and win is because their path to get to the championship isn’t as difficult as lower seeds. In other words, they have fewer tough games against better teams. 1 seeds are rewarded for their seeding so it’s a good idea to pick them to go far.

What stats are most important to winning? What are stats like of teams that won the tournament? The top 5 stats that positively contribute to winning are a teams power rating, their adjusted offensive efficiency, their effective field goal percentage, and their 2-point and 3-point shooting average. This means that offensively, teams that shoot the ball really well are in great shape for success. The top 5 stats that could hurt a team are a teams adjusted defensive efficiency, their effective field goal percentage allowed, their turnover percentage allowed, and their 2-point and 3-point shooting percentage allowed. This means that a team is in bad shape if they’re giving up a lot points, not playing good defense and allowing opposing teams to shoot higher percentages. A profile of a team that will win the tournament has averages like this in the stats mentioned above listed below:

Champion Averages

Average ADJDE: 91.4

Average ADJOE: 124.14

Average BARTHAG: 97.09

Average EFG\_O: 55.82

Average EFG\_D: 46.9

Average TOR: 15.7

Average 2P\_O: 55.16

Average 2P\_D: 46.28

Average 3P\_O: 38.0

Average 3P\_D: 31.96

Average Win%: 88.79

For people who want to pick a better bracket, my advice is to pick the best teams from ACC and Big East to go far and win the tournament. From my analysis, the top teams from these conferences have dominated the winner circle the past few years. I also understand that there are also outside factors like recruiting, coaching strategies, and injuries that may contribute to a team’s success or failures in the tournament. I didn’t include looking too deep into contributing factors like those outside of my datasets because of time I had to do the project. However, that data can be useful for putting together a more complete picture. Hopefully my data sheds some insight into how to win your March Madness leagues.