Every March, millions of sports fans watch in excitement and awe for one of the most unpredictable sports events of the year: The NCAA March Madness Tournament. First beginning in 1939 with only a measly eight teams, the tournament today features the best sixty-eight college basketball teams in the nation, each competing single-elimination style in hopes of winning the national championship. With the tournament being single elimination, there are very high stakes each and every game, allowing for plenty of room for upsets. Over the years, March Madness has earned a very well-known reputation for having dramatic upsets and Cinderella stories, which makes predicting the ultimate winner of the tournament very difficult for analysts and fans alike. This brings up many questions of curiosity, such as what metrics are typically required for a team to have a long tournament run and whether offensive or defensive efficiency has a greater impact when it comes to winning the post-season games in hopes of eventually winning it all.   
  
In hopes of predicting success in the tournament, this project aims to analyze previous NCAA basketball data to uncover the essential factors for post-season success. Using a Kaggle dataset that contains the historical data of various Division I NCAA basketball teams from 2013 to 2023, we will analyze how the regular season statistics of different teams correlate to their performance in the tournament. A linear regression model will be built, and feature analysis as well as various other statistical tests will be implemented to predict success in the tournament and to help identify the most impactful metrics that makes a team possible “championship contenders”. By understanding the factors that directly influence success in the post season, we can help fans, analysts, and coaches better comprehend what it takes to win it all.  
  
First we needed to set up our R environment with the necessary packages to preprocess our basketball data, explore various metrics, and develop models to predict tournament success. In our project, we incorporated nine different R packages to help us achieve our goal of an effective and accurate March Madness prediction model. First, we had the dplyr and tidyverse packages to assist with data manipulation and necessary cleaning. To further assist us with data manipulation we also used the reshape2 package. Next, we implemented the ggplot2 and corrplot packages to help with data visualizations, especially pertaining to our correlation matrices. We then used the caret package to help us with building our machine learning model, along with the car and MASS packages to help with further statistical testing and analysis. Finally for ease of life purposes, we included the here package to help simplify file path management, so we didn’t need to incorporate hardcoded paths throughout the project.  
  
After collecting the necessary packages for our project’s library, we then went ahead and loaded in our Kaggle dataset. We then implemented our first instance of data cleaning by removing instances where a team did not participate in the tournament. We then converted the Seed column entirely, adjusting it into a numerical format and creating a custom function that maps all possible tournament outcomes into corresponding numbers. The outcomes range from one to eight, with one corresponding to losing in the First Four (which determines the final four teams that advance to the main 64-team bracket) and eight corresponding to winning the tournament all together. We did this to ensure only relevant teams are included in our analysis and prepare our data for various modeling and visualization.  
  
Next we create visualizations to analyze various relationships between team metrics and postseason performance in the March Madness tournament. Each correlation analysis plot we created investigates the connection between how far a team advanced into the tournament (POSTSEASON\_NUMERIC) and various team metrics. The metrics from our dataset we decided to visualize were offensive efficiency (ADJOE), defensive efficiency (ADJDE), power rating (BARTHAG), effective field goal percentage made (EFG\_O), effective field goal percentage allowed (EFG\_D), wins during the regular season (W), and finally seed placement. Using various scatter plots with linear regression lines, we were able to easily detect valuable trends and evaluate which metrics had a positive correlation with postseason performance, and which metrics possibly hurt a team’s chance of preforming well.  
  
For the first scatter plot we plotted offensive efficiency (ADJOE) against postseason performance (POSTSEASON\_NUMERIC). We observed that there is a positive correlation, meaning that teams with higher offensive efficiencies tend to advance into the further rounds of the tournament, suggesting that strong offensive play is an important factor for postseason success.  
  
For the second scatter plot we plotted defensive efficiency (ADJDE) against postseason performance (POSTSEASON\_NUMERIC). We observed that there is a negative correlation, meaning that teams with lower defensive efficiencies (remember lower is better!) tend to advance into later rounds of the tournament, aligning with the common idea that defense is crucial in basketball, especially in a tournament where a single loss eliminates you.  
  
For the third scatter plot we plotted power rating (BARTHAG) against postseason performance (POSTSEASON\_NUMERIC). We observed that there is a positive correlation, meaning programs who achieve a higher BARTHAG power rating tend to advance further in the tournament, highlighting that a team’s overall quality is more important for postseason success, rather than just relying on a few key players.  
  
For the fourth scatter plot we plotted effective field goal percentage made (EFG\_O) against postseason performance (POSTSEASON\_NUMERIC). We observed that there is a positive correlation between the two variables, indicating that teams who have a better shooting efficiency tend to make deeper tournament runs. The ability to score effectively and efficiently is crucial, as a team only has so many opportunities in a game to score.  
  
For the fifth plot we analyzed effective field goal percentage allowed (EFG\_D) against postseason performance (POSTSEASON\_NUMERIC). We found that there was a negative correlation, which implies that teams who allow lower field goal percentages from opponents achieve more post-season success, reinforcing the fact that defense is key in the tournament.  
  
Our sixth scatter plot compared regular seasons wins (W) against postseason performance (POSTSEASON\_NUMERIC). We observed that there is a very strong positive correlation between the metrics, meaning that teams that achieve more wins during the regular seasons tend to perform better in the postseason, reflecting that overall team strength and consistency plays an essential role in determining tournament success.  
  
Our seventh scatter plot displays wins above bubble (WAB) against postseason performance (POSTSEASON\_NUMERIC). We observed that there is a positive correlation, meaning that teams with higher WAB values (a team is performing well relative to what is expected of an average bubble team) are more likely to be successful in the postseason, detailing that teams who outperform their original expectations tend to carry that momentum into the tournament.   
  
For the eighth and final scatter plot we plotted SEED against postseason performance (POSTSEASON\_NUMERIC). In the plot a strong negative correlation was present, which makes sense as lower seeds tend to advance further into the tournament than the higher seeds. This implies that the tournament seeding itself is a significant predictor of postseason success, as teams who receive lower seeds are typically more qualified for the championship.