NCAA Basketball Analysis

As we all know, basketball is a very popular sport and includes many different types of statistics. Whether it be a team’s shooting percentage or an individual player’s amount of playing time, just about any statistic collected in basketball can be utilized and analyzed to try to help a team obtain its best chance at success. That being said, this project provides a statistical analysis of all of the Division 1 teams in the NCAA 2020 basketball season, and tries to uncover what makes a basketball team successful.

**Data**

The data that is utilized in this statistical analysis is obtained from Kaggle.com, which contains statistics on each team’s performances in seasons over the years. It should be noted that data for the 2020 season was selected specifically for this analysis, but does not contain any postseason statistics due to the season being cut short because of COVID-19. That being said, comparing these results to previous years could prove to be inconsistent as playoff performances were included in years past, but not for 2020.

The dataset included over twenty different statistics, but only ten were selected for this analysis.The ten variables selected were deemed to provide a sufficient range of information and are as follows:

* Team
* Wins
* EFG\_0: effective field goal percentage shot
* EFG\_D: effective field goal percentage allowed
* TOR: Turnover percentage allowed (Turnover Rate)
* TORD: turnover percentage committed (Steal Rate)
* ORB: offensive rebound percentage
* DRB: defensive rebound percentage
* FTR: free throw rate (How often the given team shoots Free Throws)
* FTRD: free throw rate allowed

**Methods**

The methods that are incorporated in this analysis are principal component analysis, factor analysis, and cluster analysis. Principal component analysis was utilized in order to decide which were the most significant factors in determining the success of an NCAA basketball team. Factor analysis was used to test how these predictors of success could be grouped together and what conclusions could be drawn based on the groupings. Finally, cluster analysis was used to determine how to group these teams together in relation to one another and to test if the basketball statistics selected, really determine the overall success of each team.

When looking at the principal component analysis, it can be determined that three principal components will explain most of the variance in the data, and in fact explain about 63% of the total sample variance. Three principal components were deemed a sufficient amount of components to include because the first three components had an eigenvalue over one. Additionally, an analysis of the scree plot shows that after three components the graph starts leveling off which implies that using more than 3 principal components would not be as meaningful. It can be seen that the first principal component shows a negative value for wins, effective field goal percentage shot and offensive rebounding percentage. Since the biggest negative value is under wins, this shows the difference between the outcome of the game and how well the team plays. The second principal component has negative values for effective field goal percentage shot and effective field goal percentage allowed, which could show the difference between the teams contested shot percentage compared to how the team plays on the court. This shows how important overall success in basketball is compared to just being a proficient shooting team. The third principal component shows negative values for effective field goal percentage allowed, turnover rate allowed, offensive rebounding percentage, and free throw rate. Since the largest negative values are variables describing how well the team scrap for the ball, this component could show the difference between hustle plays and regular stats.

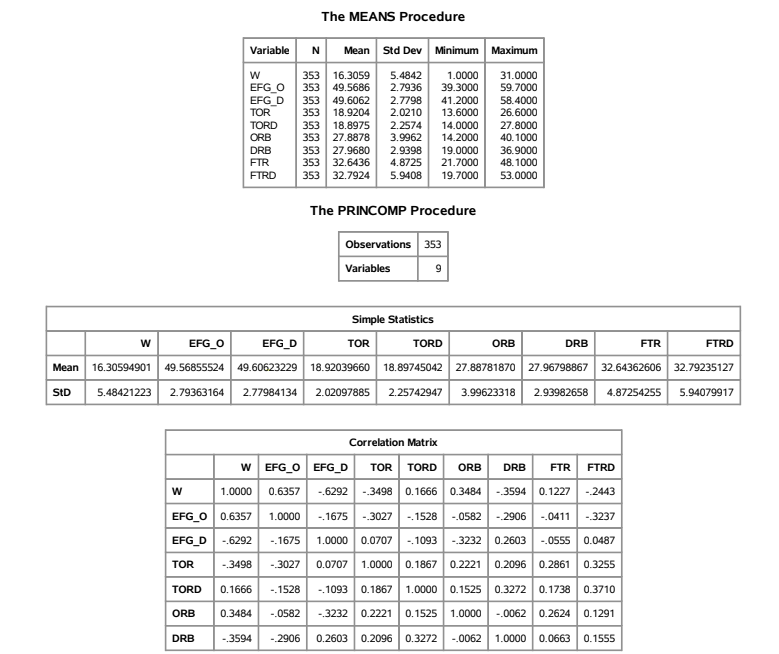
In doing a factor analysis, principal component method and maximum likelihood method were utilized. Upon initially doing the factor analysis, the factors were tricky to evaluate and the results of the ML and PC methods were inconsistent. Thus, an orthogonal rotation was utilized and by using a varimax rotation for both methods, the factors were a lot easier to interpret. It can be seen that both methods separated the factor into distinct groups with wins and effective field goal percentage having positive loadings on factor one, turnover rate, offensive rebound %, free throw rate having positive loadings on factor two, and effective field goal % allowed, steal rate, defensive rebound %, free throw rate allowed having positive factors loadings on factor three. With a little bit of background knowledge of basketball, it becomes clear that the variables that are grouped in factor two include offensive stats, and the variables included with factor three are defensive statistics. Factor one includes wins and effective field goal percentage shot, which is interesting because effective field goal percentage shot is an offensive statistic but is not grouped with the other offensive stats in factor two. This could indicate that the amount of field goals a team successfully makes is a bigger predictor of success than the other offensive statistics, which makes sense, as the team who scores more wins in basketball. This conclusion is further supported by looking at the variance explained chart and seeing that factor one(wins and effective field goal percentage shot) explains the highest amount of variance in the data.

When doing a cluster analysis, complete linkage was utilized, as using a single linkage did not prove to be that effective. The results show that all of the teams playing division NCAA basketball were separated into three separate clusters. Upon inspecting the clusters it can be seen that cluster one has the lowest mean win percentage and effective field goal shot percentage, cluster three has the highest of these two statistics, and cluster two falls somewhere in the middle of the two. Thus, it was determined that cluster three included the best teams in the league, cluster one included the worst teams in the league, and cluster two included more of the middle-tier teams. This conclusion is confirmed when looking at the whole cluster history and seeing that top teams in the league such as Gonzaga, Kansas, and San Diego State are included in cluster three.

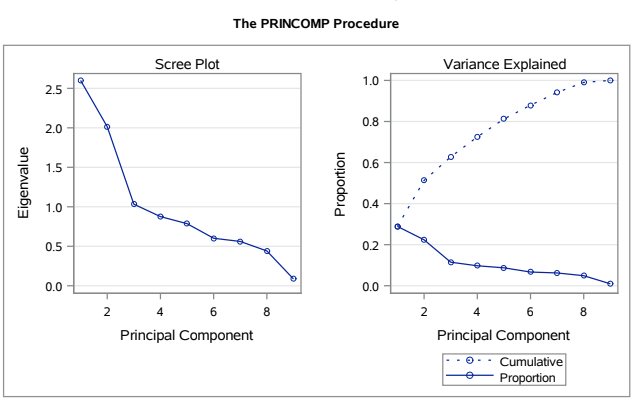
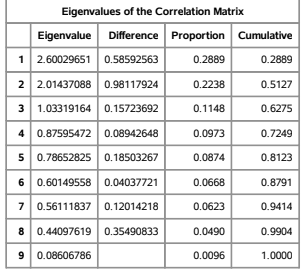
**Conclusion**

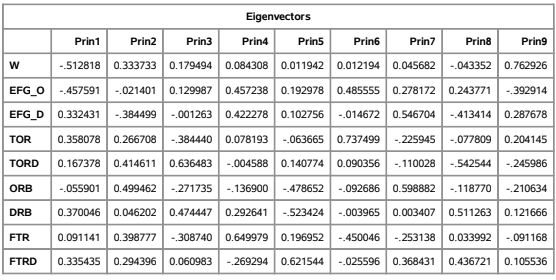
Clearly, basketball is a sport that collects a wide range of statistics, and doing statistical analyses on this data can really give insight into the different components that go into the success of a basketball team. Through the use of principal component analysis, factor analysis, and cluster analysis, we began to understand what drives a team’s success and which teams were more effective in different areas. Although success is very subjective, running tests on quantifiable data points can give a better understanding of what the success of a basketball team truly means. By analyzing different basketball stats, teams can utilize this information to see what areas they need to work on, and what areas give them the best chance of winning. In this case, we determined that effective field goal percentage of shots made was the biggest indicator of wins for a team, so this means that teams definitely need to work on their shooting extra hard in practice to maximize their teams success! Additionally, for anyone looking to make bets or predictions on future seasons, looking at patterns in these data sets, or pinpointing how well a team does in a pivotal area, can even help them make their predictions more successfully. Although these tests gave a general idea of what success means for a basketball team, further research could be done on a specific team’s data and could be analyzed to see which areas their team specifically excels in and which they need to work on.

Appendix A-1 Means Procedure, Simple Statistics, and Correlation Matrix



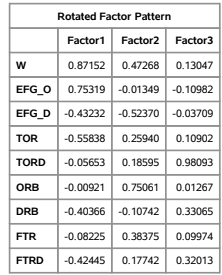
Appendix A-2 Principal Component Analysis



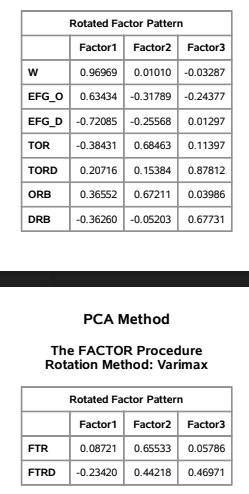
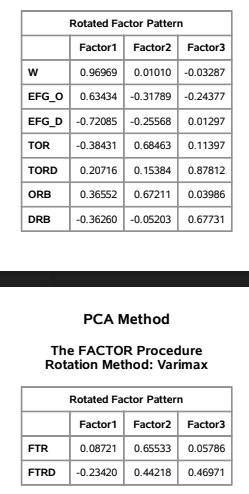


Appendix A-3 Factor Analysis

**ML Method**



**PC method**

****

Appendix A-4 Cluster Analysis

