College Football Team Roster Composition and Relationship with Team Performance

by

Ansen Gunawan

Honors Thesis Project

Appalachian State University

Submitted to the Department of Mathematical Sciences
in partial fulfillment of the requirements for the degree of
Mathematics, Bachelor of Science

May, 2020

Approved by:

Ross Gosky, Ph.D., Thesis Director, Mathematical Sciences

William J. Cook, Ph.D., Honors Director Department of
Mathematical Sciences

Eric Marland, Ph.D., Chair Department of Mathematical Sciences

Abstract:

This paper proposes and compares a set of models of college football team rosters for teams in major conferences during the years of 2018-2019. A cluster analysis was performed to classify groups of teams, using the number of scholarship players at various position groups, the class years of those players, other physical attributes such as height and weight, as well as the players' recruiting rankings on www.247sports.com. Once the clusters were determined, we examined the clusters for a relationship between the clusters and team

1. Introduction

College football is responsible for bringing in millions in revenue for schools, and it can be observed that some programs are more notable than others. Hence, the question on how are teams constructed and the impact it has on team level success are raised. A team's success is a source of significant interest, particularly for coaches in the largest conferences, such as the Southeastern Conference (SEC), the Big Ten, the Big Twelve, the Pac-12, and the Atlantic Coast Conference (ACC). With the increased popularity in college football, several media outlets publish data on team recruits annually such as www.rivals.com, www.espn.com, and www.247sports.com.

Each year, recruits across the United States will sign a letter of intent by National Signing Day to let coaches and fans know where they decide to land. Looking specifically at recruits accepting to play football at Football Bowl Subdivision (FBS) schools, these recruits are typically assigned a recruiting ranking. These star ratings are a quantitative measure to convey the talent of a specific recruit, ranging from two stars to five stars. Different media outlets will use different algorithms, and some sites will also place a numeric rating on players as well but the higher rated a player is, the better they are projected to perform on the college level.

College football teams in the (FBS) subdivision can have a total of 85 players on scholarship at any given time. Given the limited supply of scholarships, coaches take great measures in recruiting the players they want that will better their team. However, is there a particular method for teams to recruit their players? For example, one team may prefer an offensive focus with bigger receivers, offensive lines, etc. over defensive players. Furthermore, is there a particular team construction format that is more successful than other formats? These are questions coaches will ask themselves going into the recruiting season when analyzing players in terms of ratings and physical attributes.

Roster construction has been the subject of curiosity and analysis in recent years, and is not limited to college football. One study conducted by Peterson at the University of Northern Iowa for a thesis looked at the relationship between team characteristics and team defenses in the NBA. Using several measurable factors of teams affect roster construction in order to predict the team's defensive ability (Peterson, 2014). Another study on NBA's roster construction was also done combining advanced analytics and traditional evaluation methods to help general managers with strategies and tools to maximize team performances (Mills, 2015).

Many popular press articles have been written on the subject such as Boyd (2014). In this particular article, rosters are broken down into different formats. The most basic is having forty-one scholarships given to both offensive and defensive players and three scholarships allotted to kickers, punters, and long snappers. However, rosters can be broken further with defensive schemes such as 4-3 defense, 3-4 defense, etc. and this will influence recruitment since 4-3 defense requires heavier and bigger defensive players over a 3-4 defensive scheme which prefers more agile and quick defensive players. This article also looks at offensive schemes as well. For example, a spread-to-run offense will look for quick and lighter offensive skilled positions over a pro-style offense which prefers bigger and heavier skilled positions. Each system has its own requirements and targets particular types of players that will have the greatest impact for the system (Boyd, 2014).

Roster construction is important for teams' success at the college level. To measure a team's success in a given season, numerous metrics are available such as being ranked in polls such as the Associated Press (AP) or Coaches poll. These polls only rank 25 teams, and because of this, many teams are not evaluated in the polls because they do not receive votes. Another way to measure success is the winning percentage of a team given a particular season but this fails to take strength of schedule into account. One ranking system that takes into account strength of schedule is Sagarin rating. The Sagarin rating is produced by Jeff Sagarin (www.sagarin.com) and it attempts to quantify the strength of a team in a given season. The Sagarin ratings work by providing each team a numeric score, ranging from 0 to 100, using a computational formula for the season. With these Sagarin ratings, two teams can be hypothetically matched up, and the difference in two teams' Sagarin composite ratings for that season is roughly comparable to the point differential between the two teams. In other words, a team that is 5 points higher in the Sagarin rankings than another would be favored by 5 points on a neutral field.

Our study will compare a set of models of college football team rosters, specifically for scholarship players, and observe a potential format teams are recruiting players given attributes and recruiting rankings. Attributes being discerned are position of players, height and weight, class year, and recruiting ratings compiled from www.247sports.com. We then perform a cluster analysis to group teams that construct their rosters with similar formats. As a secondary objective, we examine the relationship between roster construction and team’s success.

2. Data Collection

We collected data from www.247sports.com for the years 2018 to 2020, and have specifically focused on teams in the largest conferences, specifically the SEC, Big Ten, Big Twelve, Pac-12, and ACC. These Power Five conferences were chosen as our priority since data was readily available and historically, higher rated players are recruited to play at these conferences. Furthermore, recruiting rankings tend to vary the most among teams in the major conferences.

For two seasons, the 2018-2019 season and the 2019-2020 season, we examined the roster composition of each team in the Power Five conferences. When a player's position was recorded, there was some inconsistency in position designation across the different teams. For example, some rosters used "DB" for Defensive Backs, regardless of whether the player was a Cornerback (CB) or a Safety (S). Other team rosters made this distinction. For this reason, we grouped players into position groups, which we called standard positions.

Some other summarization completed for position groups: kickers, long snappers, and punters were grouped into "ST" or special teams. Offensive lines, "OL" consists of centers, offensive linemen, and offensive guards. Wide receivers and tight ends were grouped together as "WRTE" to have a better observation at receiving cores for teams. Full backs and running backs are standardized into one group, "RB". Defensive lines, "DL", consists of defensive linemen, defensive ends, nose tackles, and defensive tackles. Defensive backs, "DB", consists of defensive backs, safeties, and cornerback. Standardizing the positions makes grouping of offensive line, defensive line, running back, receiving core, defensive backs, and special teams possible. These groupings allow a consistent listing of positions across different rosters, allowing for positional analysis of rosters to be part of our analysis.

The data set is composed of the following variables (and variable names) measured for each player for each season:

- 247sports recruiting measurement of the past two season
    - Recruiting measurement from www.247sports.com, (0 to 1), where higher numbers indicate a higher recruiting rating
    - Class year (Fr, So, Jr, Sr, and Redshirt Senior)
    - Height (in.)
    - Weight (lbs.)
    - Standard position
- Team names
- Season
    - 2018, 2019
- Conference affiliation
    - ACC, Big 10, Big 12, Pac 12, SEC

The data set is separated by season to run cluster analysis on the 2018-2019 and 2019-2020 season individually as we are determining a format of roster construction and how it affects the team's success for each given year. Additionally, we were interested in observing any significant changes throughout the two seasons. In the raw data set, some players do not have a recruiting rating on www.247sports.com. Typically this means that such players are either walk-ons or players coming out of highschools that are not viewed highly and so they are not offered scholarships. These non-scholarship players are removed from the raw data because of the focus of analysis on scholarship players. This is due to the fact that non-scholarship players are not typically key contributors to the team, and their recruiting ratings can often be set to zero if they were not heavily scouted as a high-school athlete. The data set is then arranged by teams and players are ranked starting from 1 to 85, in descending order based upon recruiting ranking, for each team. These ranks are determined by their www.247sports.com rating with 1 being the highest rated player in that team. Since this study will only be observing scholarship players, each team is filtered so that the top 85 rated players are left. These 85 players are assumed to represent the scholarship players for each team. From this data set consisting of scholarship players, we create two separate new data sets. Both still contain all players and teams, but one data set will include "Rating" and the other will not. The reason for this is to obtain a different perspective in terms of roster construction that does not include ratings for players and just on attributes of players. We took each player's measurements and created a team level roster summary, which, for each standard position, consisted of a summary of all the players on the roster at that position. We created the following summaries:

- Average recruiting rating
- Average height
- Average weight
- Number of players in each class year

Once transformed, a cluster analysis is carried out on the two data frames. Specifically, hierarchical clustering and K-Mean clustering methods will be the primary method of cluster analysis. The hierarchical clustering works by treating each observation as a separate cluster and repeatedly executes the following two steps: (1) identify pairs of clusters that are closest together, and (2) merge the two most similar clusters as one moves up the hierarchy. This particular method provides a dendrogram as a visualization of how teams are related that will be discussed later on. The second method, K-Means, works similar to hierarchical clusters, but the number of clusters can be chosen as fixed which differs from the hierarchical clustering, which allows a visualization of the clustering process with differing number of clusters. K-Means works to make the inter-cluster data points as similar as possible while also keeping the clusters as different as possible. For this paper, a K-Means cluster analysis was conducted with a minimum of two clusters up to a maximum of six clusters. We will also label each team of the clustering vector they fall in to gain a sense of which teams are clustered. As a secondary objective, we wanted to discover if there is a relationship between roster construction and teams' success. For each K-Means clustering, each team's winning percentage will be calculated and then be used to compare to other clusters. Additionally, Sagarin ratings will be incorporated and an average Sagarin rating for each cluster will also be calculated.

3. Hierarchical Clustering

After creating team level roster summaries, we performed hierarchical clustering on the teams for each season, With Hierarchical Clustering, we are offered numerous methods of clustering but only two methods will be used in this paper; complete linkage, and average linkage. Complete-linkage works by taking into account the distance between the farthest pair of observations in two clusters are measured. This linkage usually provides tighter clusters than single-linkage which takes into account the shortest distance between a pair of observations in two clusters. The second method, average linkage, sums the distance between each pair of observations in each cluster and divides by the number of pairs to get an average inter-cluster distance (Christopher D Manning, Mark Craven, Ido Dagan, et.al, 2008). Both methods provide a dendrogram for visualization of linkage and which teams are similar to one another.

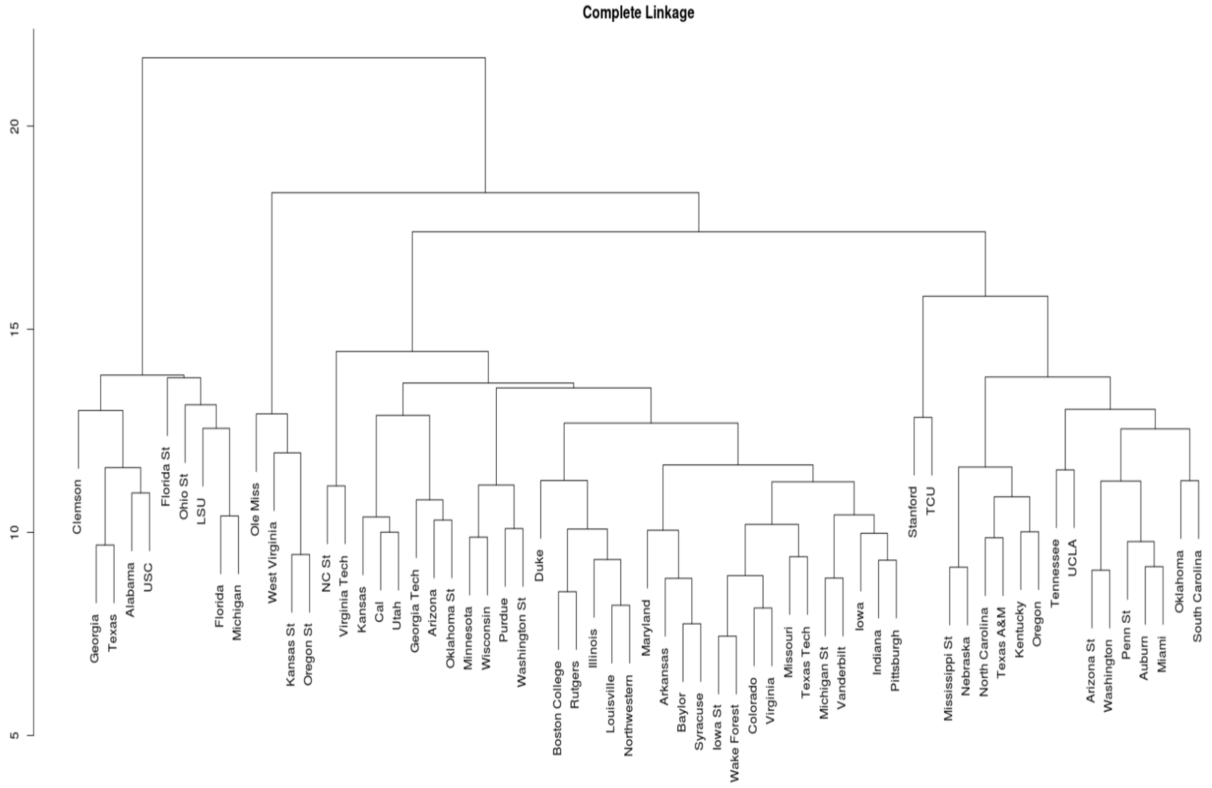


Figure 3.1: Provided is a dendrogram created using Hierarchical Clustering with Complete linkage on the “Unedited” data set of 2018.

A few findings from Figure 3.1:

- Top 3 nationally ranked teams of 2018 belong in the same cluster (Clemson, Alabama, Georgia)
- Top Power Five schools based on AP Polls are clustered together (Clemson, Georgia, Alabama, Texas, Florida State, Ohio State, LSU, Florida, and Michigan)
- Mid-level Power Five schools are clustered together and similar but further breakdown occurs as more clusters are added

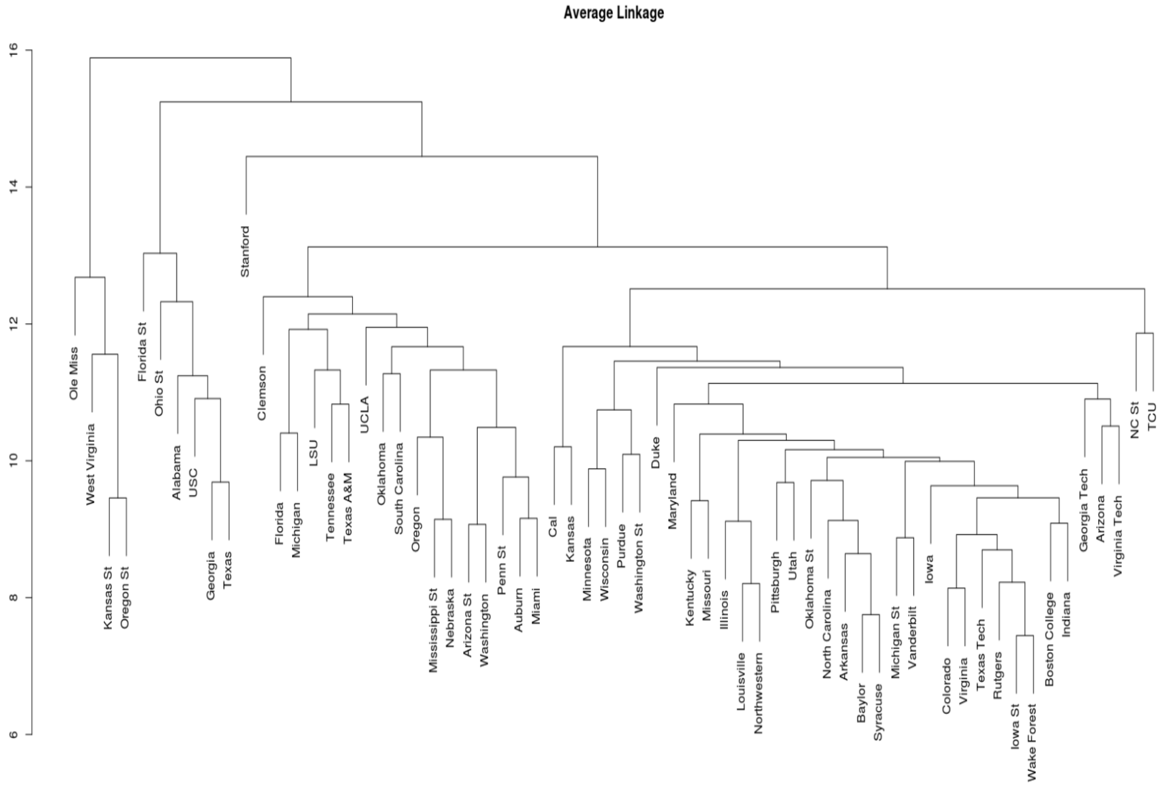


Figure 3.2: Provided is a Dendrogram using Hierarchical Clustering with Average linkage on the "Unedited" data set of 2018.

A few findings from Figure 3.2:

- Four teams are significantly different from the other teams in the data set when the first cluster split is determined (Ole Miss, West Virginia, Kansas State, and Oregon State)
- Stanford is different from the majority of teams when broken into four clusters
- Top Power Five schools are not all clustered together as one would expect

There are some differences between the clusters resulting from the complete and average linkage of the "Unedited" 2018 data set which is summarized in Figure 3.2. The complete linkage approach groups notable Power Five schools mentioned in Figure 3.1 in the same cluster whereas the average linkage breaks those particular schools into two different clusters. For example, Clemson, Michigan, and LSU are in a separate cluster when they were previously grouped together with Ohio State, Alabama, and Georgia. However, we can conclude that top Power Five schools are generally more similar in roster construction, and this is expected as more often than not, higher rated players tend to sign to these bigger schools. Another conclusion that can be made for the 2018 season is that mid-level and low-level Power Five schools are all clustered close to each other, meaning they are all similar in roster construction.

The additions of the 2019 "Unedited" data set allows us to observe if teams are clustered similarly in another season. Another observation that can be made with another season is if there are any changes between the two seasons.

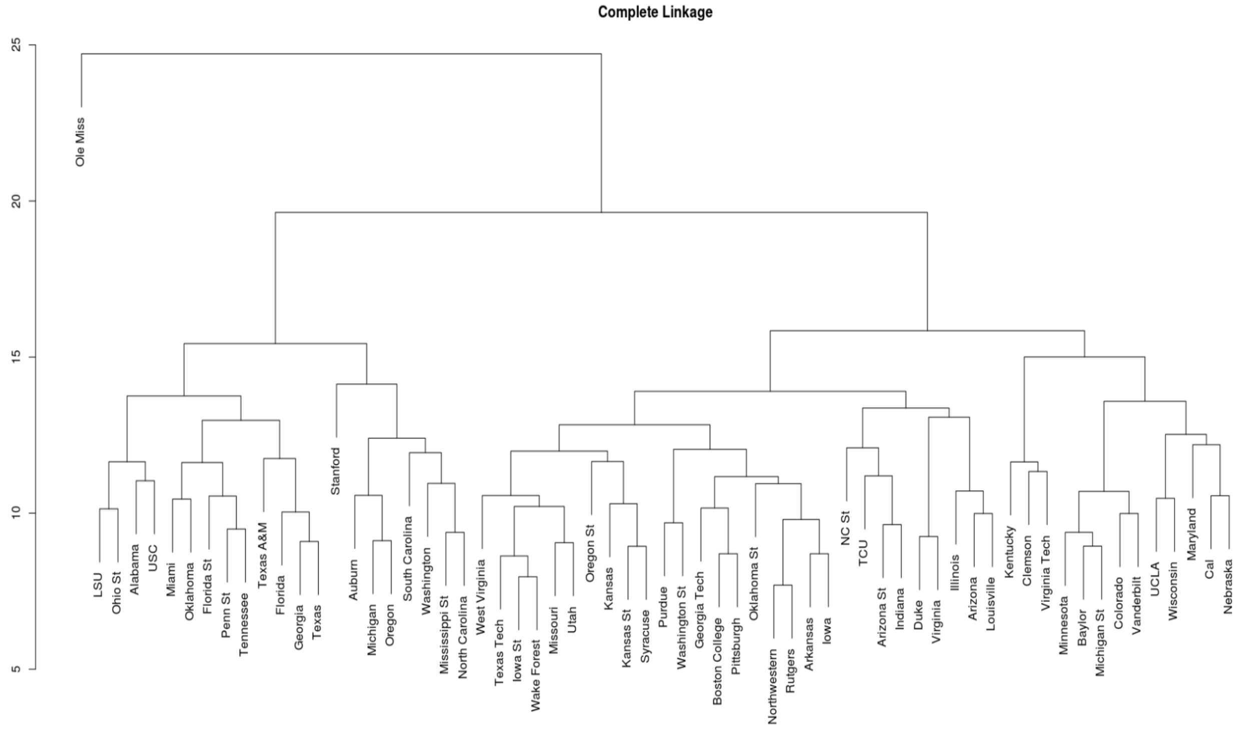


Figure 3.3: Dendrogram using Hierarchical Clustering with Complete linkage on the "Unedited" data set of 2019.

A few findings from Figure 3.3:

- Ole Miss is constructed differently from all the other teams
- Top Power Five schools are clustered together; however, Clemson does not appear to be similar to those schools clustered (LSU, Ohio State, Alabama, Oklahoma, Georgia)
- Clemson appears to be more similar to mid-level and low-level Power Five schools

Comparing the two seasons' hierarchical clustering with complete linkage, significant changes have occurred. As stated above, Clemson appears to be more similar to mid-level and low-level Power Five schools such as Kentucky, Virginia Tech, Maryland. With Wisconsin being the only high-level school closely related in the 2019 season. In the 2018 season, Clemson was clustered closely to Georgia, Alabama, LSU, and other top flight programs. Another significant change is Ole Miss clustered as an individual whereas in the 2018 season, Ole Miss was clustered closely to other lower tier programs. However, we do still note that a majority of the top programs are clustered closely which can allude to a notion that they are constructed alike. The same can be said with lower tier programs.

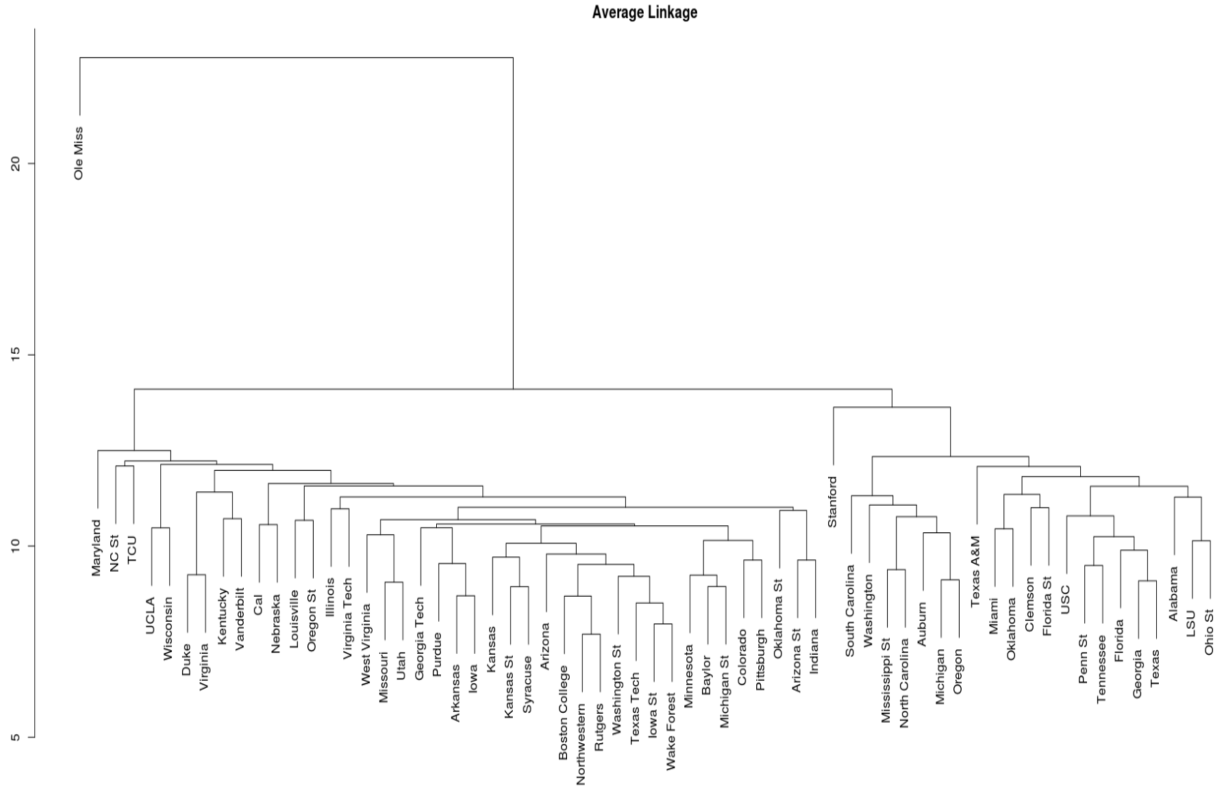


Figure 3.4: Dendrogram using Hierarchical Clustering with Average linkage on the "Unedited" data set of 2019.

A few findings from Figure 3.4:

- Ole Miss is constructed differently from all the other teams
- Top Power Five schools are clustered together (LSU, Ohio State, Alabama, Oklahoma, Georgia, Clemson)

Analogous to the dendrogram of the 2019 "Unedited" Complete linkage, Ole Miss is again clustered separately from other schools. However, there are some differences in the 2019 hierarchical cluster of both complete and average linkage. The complete linkage approach clustered Clemson and Wisconsin in a separate cluster from other top Power Five programs whereas the average linkage cluster included in the same Clemson with those same top programs. Despite this issue, we will delve deeper into the clusters provided by the Average linkage of the two seasons through K-means clustering. K-means clustering works comparably to Average linkage by averaging the distance between a pair of observations. It then provides a chart of metrics of each cluster where we can observe the differences numerically.

4. K-Means Clustering

We performed K-Means clustering as our second method of clustering on the teams for each season, one including the recruiting rankings (Unedited) and the other without them (Edited). As described previously, K-Means works similar to hierarchical clusters but with a predetermined number of clusters that are specified in the analysis. From reviewing our hierarchical cluster results, we chose 5 clusters for this next procedure as a good number to observe how teams are clustered. Referring back to Figure 3.1, for example, we see that 5 clusters incorporate large breakages between teams and avoids the issue of clustering with possible indiscernible differences. We will then examine the clusters identified in the K-Means procedures for their most notable differences.

	STmeanrating	STmeanWeight	DLmeanrating	DLmeanWeight	OLmeanrating
1	0.7997565	188.5833	0.8535731	278.1333	0.8510858
2	0.8031556	210.4815	0.8923874	297.4161	0.8861244
3	0	0	0.8447676	278.2382	0.8471527
4	0.7952673	210.9444	0.8551251	274.9129	0.8534111
5	0.8009191	199.4424	0.8646885	277.0074	0.8625156

	OLmeanWeight	WRTEmeanrating	WRTEmeanWeight	RBmeanrating	RBmeanWeight
1	302.3046	0.8528018	209.9395	0.8589543	203.8156
2	315.314	0.8872251	219.825	0.8939528	210.763
3	307.4398	0.8570163	208.3509	0.8511	210.5
4	307.1381	0.8597776	207.2622	0.857443	209.0563
5	304.6989	0.8706029	235.3415	0.8689813	214.0613

Figure 4.1: Sample chart created from K-Means Clustering on “Unedited” data set of 2018.

A few findings Figure 4.1:

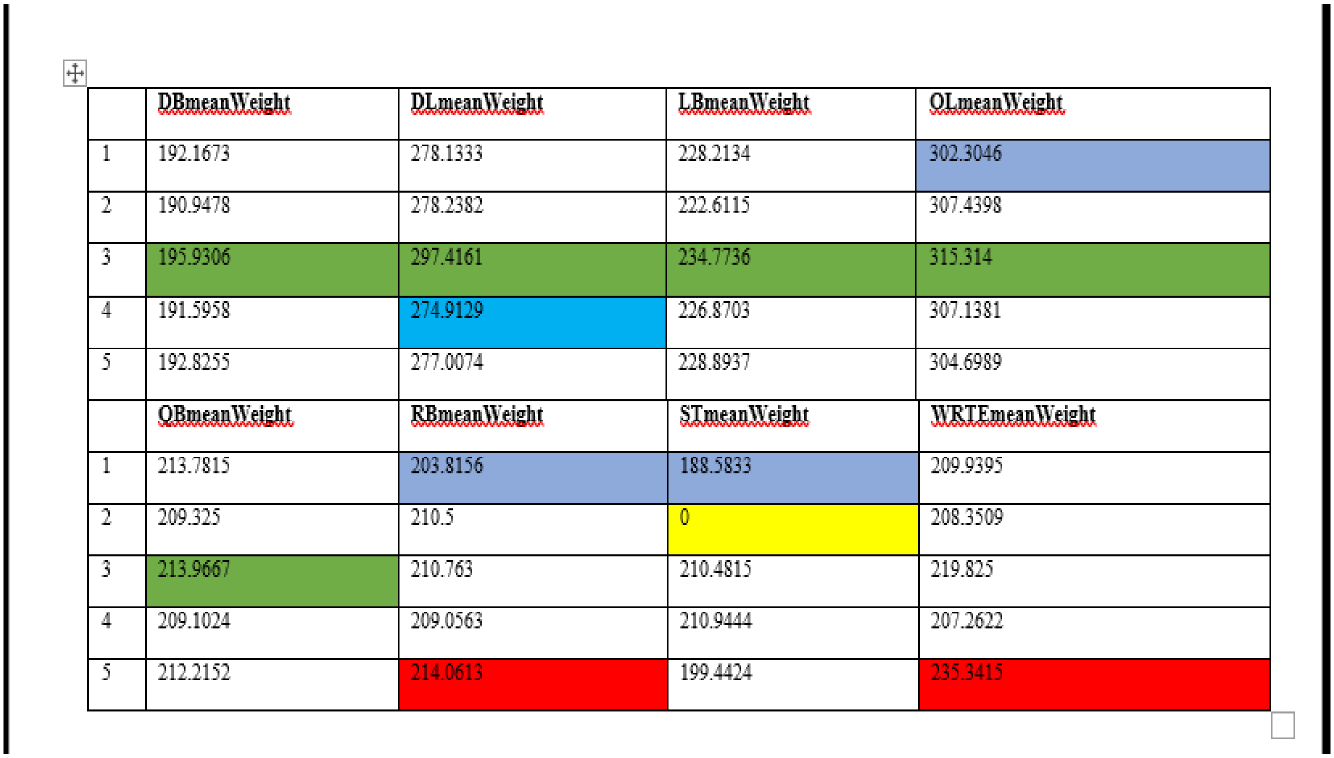
- Cluster 3 contains no special teams data
- Cluster 1 appears to have the lightest in terms of mean weight of OL, ST, and RB
- Cluster 2 appears to have the highest position mean rating and heaviest defensive position as well as OL
- Cluster 4 has the lightest DL in mean weight
- Cluster 5 has the heaviest RB and WRTE in mean weight

Cluster:	Team
1	Arizona, Penn State, Tennessee, Iowa
2	Alabama, Georgia, LSU, USC
3	Kansas State, Ole Miss, Oregon State, West Virginia
4	Miami, Oklahoma, Texas, Wisconsin, Ohio State
5	South Carolina, Clemson, Duke, Michigan, NC State

Figure 4.2: Sample of teams in each cluster of K-Means on “Unedited” data set of 2018.

As with the hierarchical clustering analysis, K-Means analysis was done on each data set to discern the differences in clusters. In Figure 4.2, we are able to get a snapshot of cluster identification for 2018 “Unedited” which as a reminder includes players’ ratings. From this Figure 4.2, we recognize that big Power Five programs such as Clemson and Michigan are clustered differently than Alabama, Georgia, and LSU. The Average Linkage of the hierarchical clustering of 2018 preludes a similar breaking in the top Power Five programs between clusters that is seen in K-Means. Furthermore, the Average Linkages also clustered Kansas State, Ole Miss, Oregon State, and West Virginia together and with K-Means analysis, those four teams appear to have no scholarship players that are designated as Special Teams. Investigating further, Average Linkage clusters of hierarchical clustering are noticeably similar to the clusters provided by K-Means analysis. With this in mind, the results given by K-Means provide insight in the differences between the clusters. Referring to Figure 4.1, we see that cluster 2 contains the highest position mean and the heaviest defensive players and offensive lines. In this cluster. Alabama, Georgia, and LSU stand out, and with these Power Five programs grouped together, it is expected this cluster will recruit the highest rated players available. The likely relationship with ratings and size metrics of these players probably explains why this cluster also has the heaviest players described. Cluster 5, which contains Clemson and Michigan, two prominent teams, has the heaviest running back and receivers in weight. This may allude to how teams are built depending on the offense run at those programs. Cluster 4 has the lightest defensive lineman mean weight and this cluster is highlighted with some big and mid level programs such as Ohio State, Wisconsin, and Oklahoma. Similarly to cluster 5, perhaps programs such as these require lighter defensive lineman for scheme purposes. Lastly, cluster 1 seems to contain the lightest mean weight of offensive line, special teams and running back and consists of teams such as Penn State, Tennessee, and Arizona.

Pertaining to 2018 “Unedited” data set and summarized in Figure 4.1, it appears as if mean ratings are not significantly different between each cluster and body metrics seems to be weighted more in clustering. Additionally, Power Five programs one would expect to be clustered together are not such as Clemson being clustered separately from their peers. A K-means analysis was also done on 2018 “Edited” data with results summarized in Figure 4.3, and investigated further if mean weights is a strong factor in clustering.



	DBmeanWeight	DLmeanWeight	LBmeanWeight	OLmeanWeight
1	192.1673	278.1333	228.2134	302.3046
2	190.9478	278.2382	222.6115	307.4398
3	195.9306	297.4161	234.7736	315.314
4	191.5958	274.9129	226.8703	307.1381
5	192.8255	277.0074	228.8937	304.6989

	QBmeanWeight	RBmeanWeight	STmeanWeight	WRTEmeanWeight
1	213.7815	203.8156	188.5833	209.9395
2	209.325	210.5	0	208.3509
3	213.9667	210.763	210.4815	219.825
4	209.1024	209.0563	210.9444	207.2622
5	212.2152	214.0613	199.4424	235.3415

Figure 4.3: Sample chart created from K-Means Clustering on “Edited” data set of 2018.

A few findings from this chart:

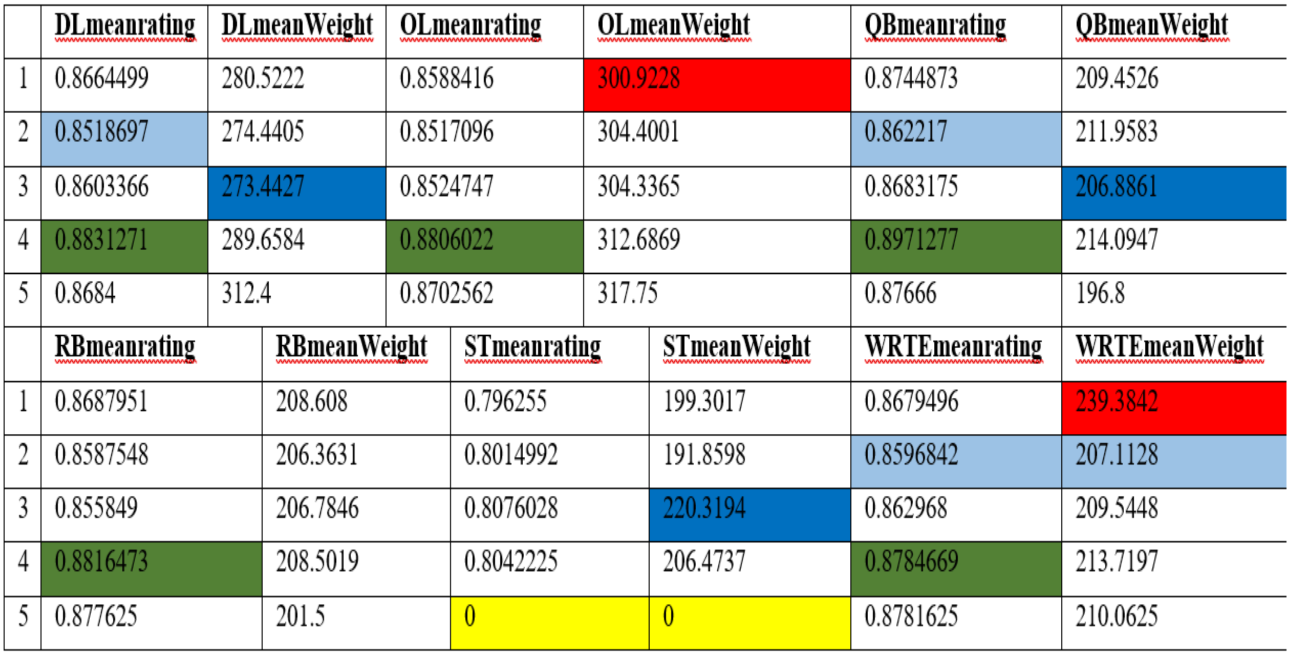
- Cluster 2 contains no special team data
- Cluster 1 appears to have the lightest in terms of mean weight of OL, ST, and RB but 2nd heaviest QB
- Cluster 3 generally has the heaviest players
- Cluster 4 has the lightest DL in mean weight
- Cluster 5 has the heaviest RB and WRTE in mean weight

Cluster:	Team
1	Arizona, Penn State, Tennessee, Iowa
2	Kansas State, Ole Miss, Oregon State, West Virginia
3	Auburn, Alabama, Georgia, LSU
4	Miami, Oklahoma, Texas, Wisconsin, Ohio State
5	South Carolina, Clemson, Duke, Michigan, NC State

Figure 4.4: Sample of teams in each cluster of K-Means on “Edited” data set of 2018.

Removing rankings out of the analysis, we generally received the same clusters with rankings. Referring to Figure 4.4, we see that some changes occured with labels but teams that were clustered together in the "Unedited" analysis are still grouped. As in the previous procedure with rankings, teams that do not allocate scholarships for special teams are still clustered together. Alabama, Georgia, and LSU which were clustered together and contained the highest position mean in the last cluster analysis are still categorized together with the heaviest mean weight of most positions. This trend of clusters being classified similarly to when players' ratings were included as more evident when cluster 4 and 5 are still being classified for having the lightest defensive lineman and running back and receivers, respectively.

In the case of 2018, players' ratings are not significant differentiating variables between clusters when compared with some other variables. Presumably, teams are constructed in terms of their play style and recruiting will be determined by the recommended stature of the athletes necessary for those roles. Another thing to note is that some less prominent Power Five programs were clustered with bigger programs. For example, Duke and NC State were clustered with Clemson and Michigan. This may further support the speculations that players' ratings are not significant factors in roster construction. However, we will observe if 2019 will provide similar results.



	DLmeanrating	DLmeanWeight	OLmeanrating	OLmeanWeight	QBmeanrating	QBmeanWeight
1	0.8664499	280.5222	0.8588416	300.9228	0.8744873	209.4526
2	0.8518697	274.4405	0.8517096	304.4001	0.862217	211.9583
3	0.8603366	273.4427	0.8524747	304.3365	0.8683175	206.8861
4	0.8831271	289.6584	0.8806022	312.6869	0.8971277	214.0947
5	0.8684	312.4	0.8702562	317.75	0.87666	196.8

	RBmeanrating	RBmeanWeight	STmeanrating	STmeanWeight	WRTEmeanrating	WRTEmeanWeight
1	0.8687951	208.608	0.796255	199.3017	0.8679496	239.3842
2	0.8587548	206.3631	0.8014992	191.8598	0.8596842	207.1128
3	0.855849	206.7846	0.8076028	220.3194	0.862968	209.5448
4	0.8816473	208.5019	0.8042225	206.4737	0.8784669	213.7197
5	0.877625	201.5	0	0	0.8781625	210.0625

Figure 4.5: Sample chart created from K-Means Clustering on the "Unedited" data set of 2019.

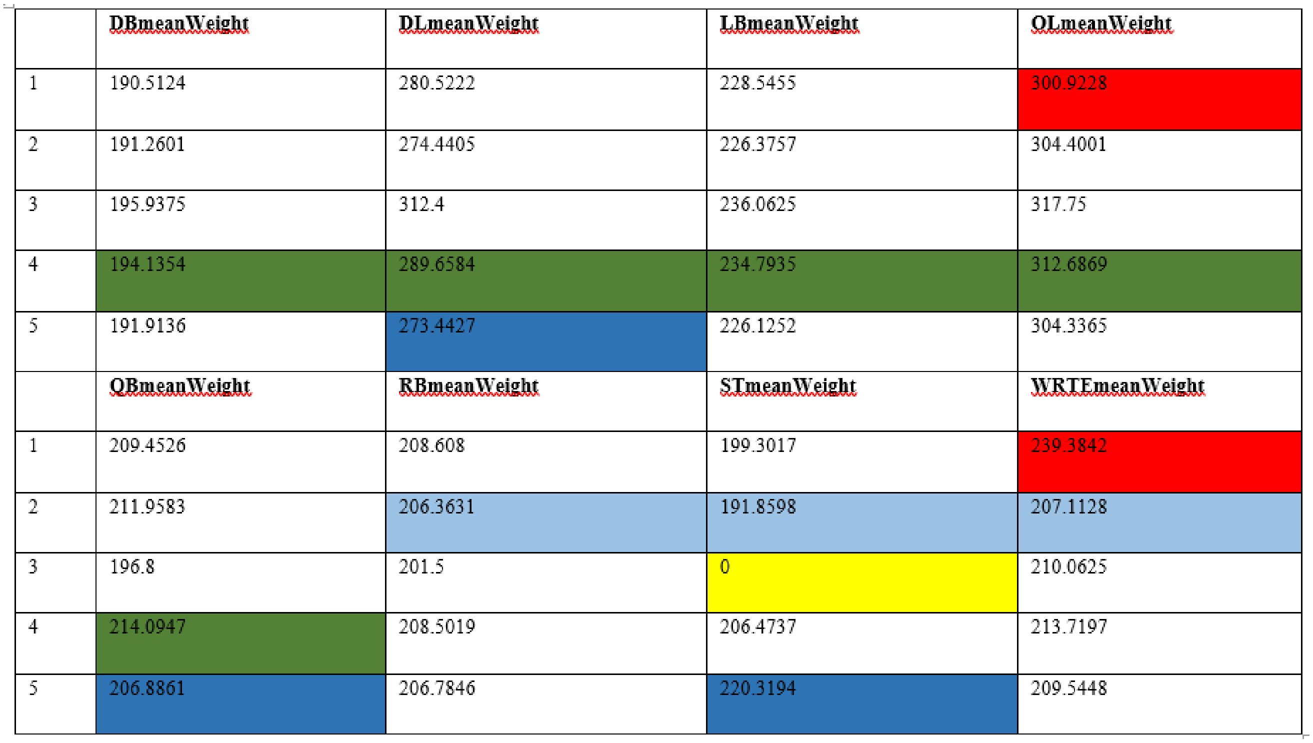
A few findings from Figure 4.5:

- Cluster 1 has the heaviest receivers, slight higher weight on RB, and lightest OL
- Cluster 2 appears to have the lowest ratings of positions and lightest receivers
- Cluster 3 appears to have the lightest QB and DL. but heaviest ST
- Cluster 4 has the highest mean rating of positions
- Cluster 5 has no ST data

Cluster:	Team
1	Clemson, Duke, Michigan State, NC State
2	Arizona State, Penn State, Baylor, Oklahoma, Texas Tech
3	Iowa, California, Georgia Tech, Missouri
4	Auburn, UNC, Alabama, Georgia, LSU
5	Ole Miss

Figure 4.6: Sample of teams in each cluster of K-Means on “Unedited” data set of 2019.

We can observe the changes in clustering of teams across the two years. In this case, we will be comparing “Unedited” data sets to each other as well as “Edited” data sets. Referring to Figure 4.6, the biggest differences between the two seasons is that Ole Miss is clustered alone in 2019, given it still does not allocate scholarships to special team players, whereas the previous three teams that were clustered with Ole Miss have migrated to other clusters. For this reason, when comparing clusters, Ole Miss will not be included in comparison between clusters since it is a single team rather than an average. In general, clusters are similar to 2018 clusters with cluster 4 being categorized as the highest mean rating of positions with Alabama, Georgia, and LSU once again highlighting this cluster. However, it appears that UNC is also categorized with these Power Five programs. Another cluster that seems to have carried over into the 2019 season is Clemson, Duke, NC State. Again with this particular cluster, the mean weight of running back and receivers is the primary distinction in this cluster along with having the lightest offensive line. Cluster 2 and cluster 3 teams from 2018 saw some changes in classification and teams included in those clusters. In 2018, teams such as Arizona State, Penn State, and Iowa were once clustered together for having the lightest mean weight of offensive line, special teams and running backs are now classified with having the lowest mean ratings of positions and lightest receivers. Furthermore, Iowa has moved to a different cluster. We also observe changes to the same cluster from 2018 that was classified with having the lightest defensive line which are now highlighted with mid to low level Power Five schools such as Iowa, California, and Missouri whereas before, Oklahoma, Wisconsin, and Ohio State highlight the particular classification. Referring to Figure 3.3 or Figure 3.4, we can note that the dendrogram alludes to these same teams being clustered as Ole Miss is clustered individually and there are some similarities between the dendrogram and k-means cluster. Leading to similar conclusions that perhaps for 2019, weight plays a more significant role than ratings.



	DBmeanWeight	DLmeanWeight	LBmeanWeight	OLmeanWeight
1	190.5124	280.5222	228.5455	300.9228
2	191.2601	274.4405	226.3757	304.4001
3	195.9375	312.4	236.0625	317.75
4	194.1354	289.6584	234.7935	312.6869
5	191.9136	273.4427	226.1252	304.3365
	QBmeanWeight	RBmeanWeight	STmeanWeight	WRTEmeanWeight
1	209.4526	208.608	199.3017	239.3842
2	211.9583	206.3631	191.8598	207.1128
3	196.8	201.5	0	210.0625
4	214.0947	208.5019	206.4737	213.7197
5	206.8861	206.7846	220.3194	209.5448

Figure 4.7: Sample chart created from K-Means Clustering on “Edited” data set of 2019.

A few findings from Figure 4.7:

- Cluster 1 has the heaviest receivers, slight higher weight on RB, and lightest OL
- Cluster 2 appears to have the lowest RB, ST, and WRTE
- Cluster 3 has no ST data
- Cluster 4 has generally the heaviest players
- Cluster 5 has heaviest ST, and lightest DL and QB

Cluster:	Team
1	Clemson, Duke, Michigan State, NC State
2	Arizona State, Penn State, Baylor, Oklahoma, Texas Tech
3	Ole Miss
4	Auburn, UNC, Alabama, Georgia, LSU
5	Iowa, California, Georgia Tech

Figure 4.8: Sample of teams in each cluster of K-Means on “Edited” data set of 2019.

Compared to 2018 “Unedited” and “Edited” comparison, the 2019 season datasets provide results that support the conclusion that weight is the determining factor in clustering of teams. Referring to Figure 4.8 and Figure 4.9, team clustering is the exact same between “Unedited” and “Edited” with a slight change in cluster designations as Ole Miss is now in cluster 3 instead of cluster 5. Furthermore, these clustering are still falling under the same classifications from “Unedited” 2019 data sets. Teams with the heaviest receivers and running back are still clustered as well as teams with the heaviest players. This may allude to a possible explanation into why some less prominent Power Five schools are grouped with higher level Power Five schools such as UNC being grouped with Auburn and Alabama: UNC recruits players of similar stature.

Another trend that appears is that generally teams clustered in 2018 are still clustered together in 2019 with the same classifications. For example, the cluster classified with the heaviest receivers and running back in 2018 are still clustered together in 2019 with some additions and losses but generally the same. Further supports the outcome being observed with a player's size being the dictating variable in clustering.

5. Cluster's Success

As a secondary objective, we observe teams' success between clusters using both the team's winning and average Sagarin rating of clusters. Following the conclusion of size being a significant variable in clustering, we will be focusing on each season's "Unedited" quantitative measures of team's success for analysis but will provide `Edited` results for reference and comparison.

2018 "Unedited"

Cluster:	Win Percentage (%)	Average Sagarin Rating
1	.5	73.36
2	.68	86.04
3	.56	71.17
4	.56	76.96
5	.63	78.80

2018 "Edited"

Cluster	Win Percentage (%)	Average Sagarin Rating
1	.502	73.53
2	.50	75.66
3	.70	85.546
4	.56	71.17
5	.62	79.36

Figure 5.1: Provided is a side by side chart of team's success metrics clustered by K-Means

2018 "Unedited"

Cluster:	Team
1	Arizona, Penn State, Tennessee, Iowa
2	Alabama, Georgia, LSU, USC
3	Kansas State, Ole Miss, Oregon State, West Virginia
4	Miami, Oklahoma, Texas, Wisconsin, Ohio State
5	South Carolina, Clemson, Duke, Michigan, NC State

2018 "Edited"

Cluster:	Team
1	Arizona, Penn State, Tennessee, Iowa
2	Miami, Oklahoma, Texas, Wisconsin, Ohio State
3	Alabama, Georgia, LSU, USC
4	Kansas State, Ole Miss, Oregon State, West Virginia
5	South Carolina, Clemson, Duke, Michigan, NC State

Figure 5.2: Provided is a sample of teams in each cluster of K-Means

It is careful to note for analysis purposes that cluster designations changed for some clusters as it did previously in K-Means between data sets of the same year. Regardless, we still observe the same clusters mentioned in the preceding section. Referring to Figure 5.1, we can see that there are small changes between "Unedited" and "Edited" data sets which further supports the conclusion mentioned in the last section. However, we can comment that cluster 2 of "Unedited" and cluster 3 of "Edited" have the highest win percentage and average Sagarin rating. These clusters are highlighted by programs such as Alabama, Georgia, and LSU. As in the last section, this particular cluster had the highest position mean rating and the majority of the heaviest position mean weights. Similarly, cluster 5 of both the "Unedited" and "Edited" analyses had the heaviest receivers and running backs of 2018 and came in second in terms of win percentage and average Sagarin rating. These two clusters are dominated by big Power Five programs and it should be expected they would reign on top for teams' success metrics. Likewise with less notable Power Five Programs clustered such as cluster 1, they reside on the bottom of these metrics. Cluster 1 was previously classified with having the lightest mean weight of offensive line, special teams and running back.

We cannot determine a correlation between weight and teams' success metrics. However, we can confidently allude to a possibility that less prominent Power Five programs, who are unable to recruit highly touted talent, are not as successful as teams that are able to recruit those highly rated players.

As we did with cluster analysis, the 2019 teams' success metrics will also be analyzed and the results are presented in Figure 5.3.

2019 "Unedited"

Cluster:	Win Percentage (%)	Average Sagarin Rating
1	.52	74.56
2	.52	73.39
3	.55	75.50
4	.67	84.45
5	.33	73.26

2019 "Edited"

Cluster	Win Percentage (%)	Average Sagarin Rating
1	.58	73.26
2	.53	74.56
3	.33	85.45
4	.67	84.45
5	.48	72.81

Figure 5.3: Team success metrics clustered by K-Means

2019 "Unedited"

Cluster:	Team
1	Arizona State, Penn State, Baylor, Oklahoma, Texas Tech
2	Iowa, California, Georgia Tech, Missouri
3	Clemson, Duke, Michigan State, NC State
4	Auburn, UNC, Alabama, Georgia, LSU
5	Ole Miss

2019 "Edited"

Cluster:	Team
1	South Carolina, Clemson, Duke, Michigan, NC State
2	Arizona State, Penn State, Baylor, Oklahoma, Texas Tech
3	Ole Miss
4	Auburn, UNC, Alabama, Georgia, LSU
5	Iowa, California, Georgia Tech, Missouri

Figure 5.4: Sample of teams in each cluster of K-Means

Again, there is a slight change in cluster designations but despite the differences, clusters are the same between K-Means and teams' success metrics. In combination of Figure 5.3 and Figure 5.4, we can note that cluster 4, highlighted by Auburn, Alabama, Georgia, and LSU, dominates in terms of win percentage and average Sagarin rating by a large margin. This particular cluster was classified with having the highest mean rating of position and a majority of the heaviest positions. Similarly, cluster 3, highlighted by Clemson and Michigan State, and classified with similar classifications from 2018 is second behind cluster 4 for win percentage and average Sagarin rating. Despite the likeness observed in clusters and teams' success metrics, 2019 sees a larger discrepancy between cluster 4 and the other clusters in win percentage and average Sagarin rating. This could be that less notable Power Five programs are clustered in with Clemson like Duke and NC State that are not as successful or highly touted in division 1 football. Additionally, cluster 1 and 2, respectively, is a mixture of big and small programs (Penn State, Oklahoma, and Arizona State).

As concluded from the 2018 data set, we can not say with confidence regarding a correlation between weight and teams' success. However, with the combination of analysis done on clustering and teams' success, we can allude to a possibility that weight is a more determining variable than players' rating since some clusters are a mixture of low and high level Power Five programs which affects the overall cluster success when comparing to each other.

6. Conclusion

Based on the hierarchical clusters (Figure 3.1 - Figure 3.4), we were able to observe a dendrogram that offered a representation of how teams were connected. Additionally, the dendrograms provided different numbers of possible clusters in the roots of each branch of the dendrogram that influenced the number of clusters that would be sufficient for K-Means cluster analysis. However, the dendrogram of the hierarchical cluster analysis did not provide any descriptions on what made teams similar or different in each branch.

In combination of K-Means analysis and the decision to choose five clusters to be sufficient from hierarchical cluster analysis, we were able to make the following conclusions for the two seasons analyzed:

- Players' ratings does not appear to be significant in clustering:
    - There appeared to be little changes in cluster between "Unedited", which included ratings, and "Edited", which did not.
    - Mean weight emerged as a determinant variable for classifications of clusters.
    - Less notable Power Five Programs were clustered with more notable Power Five Programs and further supports the conclusion that ratings is not prioritized in clustering.

- Special Team scholarship allocation was a notable difference in clustering

As a secondary objective, comparisons of teams' success metric, such as win percentage and average Sagarin rating, was conducted to observe if there was a relationship between roster construction and success. We achieved the following conclusions:

- Roster construction does appear to have some relationship with teams' success:
    - Clusters that were classified with having the highest mean position ratings and heaviest players for weight of positions (Alabama, Georgia, LSU) dominated by a large margin in win percentage and average Sagarin rating.
    - Clusters that were classified with having the lightest players for weights of positions generally came in last in win percentage and average Sagarin rating.

To conclude, our findings suggest that there are some clear differences in teams' roster constructions across the two seasons. From our cluster analyses, we were able to observe special team allocation of scholarships as the first difference in cluster breakdowns. Afterward, mean weight of positions was the defining factor in how clusters were classified rather than position ratings. It may be safe to presume that the more athletically built the players are, in this context the weight of players per positions, the higher rated they are in recruitment assessment. As so, heavier players or more athletically built players are clustered and recruited by big Power Five programs as we observed this trend in the cluster analysis. This result of clusters being seperated by mean weight would also suggest why some less well known Power Five programs are clustered with more highly notable programs as weight can be a good and a bad thing since weight factors in muscle density and fat. Despite this discrepancy, top programs were clustered together, and less well known programs were grouped together, generally, with some mixture in other clusters.

In combination of teams' success metrics, we were also able to observe a relationship between roster construction and how successful they were. However, it is important to note, we cannot argue for causation but only communicate results observed. As mentioned, a majority of top Power Five teams were clustered and it was observed that this particular cluster dominated win percentage and average Sagarin rating when comparing to other clusters. As a result, we can conclude teams who are constructed with highly touted athletes will be more successful.

However, our study has its flaws. With only two seasons of data, we will not be able to observe if this trend of clusters being defined by weight carries in past seasons. Furthermore, other variables could be considered besides physical attributes and ratings observed in this study such as speed, vertical and lateral jump, and other athletic measures that could be looked at. In a future study, more seasons and being more inclusives to other metrics will have to be taken into account to provide an accurate representation of roster construction,

References:

- Boyd, Ian. “How to Build a College Football Roster.” *SBNation.com*, SBNation.com, 14 Feb. 2014, www.sbnation.com/college-football/2014/2/14/5354962/college-football-offenses-defenses-recruiting-players.

- Christopher D Manning, Mark Craven, Ido Dagan, et al. “Introduction to Information Retrieval .” *Single-Link, Complete-Link & Average-Link Clustering*, 2008, nlp.stanford.edu/IR-book/completelink.html.

- Mills, Johnathan, “Decision-Making in the National Basketball Association: The Interaction of Advanced Analytics and Traditional Evaluation Methods” (2015). Lundquist College of Business. Honors Program Thesis.

- Peterson, Luke Ronald, "In defense of defense: a statistical look at roster construction, coaching strategy, and team defense in the National Basketball Association" (2014). University of Northern Iowa. Honors Program Theses. 100. https://scholarworks.uni.edu/hpt/100

Appendix A

R Version 3.5.2

2018 Season Analysis

~~~
```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE, results = 'hide')
library(readxl)
library(dplyr)
library(tidyr)
```
```{r importdata}
Master_Team_Roster <- read_excel("Master Team Roster Edited 2018.xlsx")
Master_Team_Roster <- na.omit(Master_Team_Roster)
Master_Team_Roster$Rating <- as.numeric(Master_Team_Roster$Rating)
Master_Team_Roster <- Master_Team_Roster %>% filter(Rating > 0)
```
```{r roster_rank}
Scholarship_Master_Team <- Master_Team_Roster %>% arrange(Team, Rating) %>%
  rename(std_position = 'Standard Position') %>%
  group_by(Team) %>%
  mutate(ranks = rank(-Rating, ties.method = "min"))
# Remove players above 85th
Scholarship_Master_Team %>% filter(ranks <= 85)
~~~

~~~
Scholarship_Master_Team <- Scholarship_Master_Team[
which(Scholarship_Master_Team$ranks <= 85), ]
```
```{r split_roster}
Scholarship_Master_Team_Pos <- Scholarship_Master_Team %>%
  group_by_at(vars(Team, std_position)) %>%
  summarise(meanrating = mean(Rating), sdrating = sd(Rating), minrating = min(Rating),
maxrating = max(Rating), medianrating = median(Rating),meanHeight = mean(Height),
meanWeight = mean(Weight)) %>%
  arrange(Team, std_position)

Scholarship_Master_Team_Pos1 <- Scholarship_Master_Team %>%
  group_by_at(vars(Team, std_position)) %>%
  summarise(meanrating = mean(Rating), medianrating = median(Rating),meanHeight =
mean(Height), meanWeight = mean(Weight)) %>%
  arrange(Team, std_position)

Scholarship_Master_Team_Pos2 <- Scholarship_Master_Team %>%
  group_by_at(vars(Team, std_position)) %>%
  summarise(meanHeight = mean(Height), meanWeight = mean(Weight)) %>%
  arrange(Team, std_position)

Scholarship_Master_Team_Yr <- Scholarship_Master_Team %>%
~~~

~~~~
  group_by_at(vars(Team, Yr, std_position)) %>%
  summarise(n = n()) %>%
  arrange(Team, Yr, std_position)
```
```{r Pos_wide}
test1 <- Scholarship_Master_Team_Pos %>%
  gather(key = statistic, value = value, c(meanrating:meanWeight))
test2 <- test1 %>%
  mutate(stpos_stat = paste0(std_position, statistic)) %>%
  select(Team, stpos_stat, value)
Scholarship_Master_Team_Pos_Wide <- test2 %>% spread(stpos_stat, value)
test11 <- Scholarship_Master_Team_Pos1 %>%
  gather(key = statistic, value = value, c(meanrating:meanWeight))
test111 <- Scholarship_Master_Team_Pos2 %>%
  gather(key = statistic, value = value, c(meanHeight:meanWeight))
test22 <- test11 %>%
  mutate(stpos_stat = paste0(std_position, statistic)) %>%
  select(Team, stpos_stat, value)
test222 <- test111 %>%
  mutate(stpos_stat = paste0(std_position, statistic)) %>%
  select(Team, stpos_stat, value)


Scholarship_Master_Team_Pos_Wide1 <- test22 %>% spread(stpos_stat, value)
~~~~

~~~
Scholarship_Master_Team_Pos_Wide2 <- test222 %>% spread(stpos_stat, value)
```
```{r Yr_wide}
test5 <- Scholarship_Master_Team_Yr %>%
  gather(key = statistic, value = value, c(n))


test6 <- test5 %>%
  mutate(stpos_n = paste0(std_position, statistic)) %>%
  select(Team, stpos_n, value)


test7 <-  test6 %>%
  mutate(stpos_nyr = paste0(stpos_n, Yr)) %>%
  ungroup() %>%
  select(Team, stpos_nyr, value)
Scholarship_Master_Team_Yr_Wide <- test7 %>% spread(stpos_nyr, value)
```
```{r merge_wide}
Merge_Scholarship_Master_Team_Master_Wide <-
merge(Scholarship_Master_Team_Pos_Wide, Scholarship_Master_Team_Yr_Wide, by =
c("Team"))
Merge_Scholarship_Master_Team_Master_Wide[is.na(Merge_Scholarship_Master_Team_Master_Wide)] <- 0
~~~

~~~
Merge_Scholarship_Master_Team_Master_Wide1 <- merge(Scholarship_Master_Team_Pos_Wide1, Scholarship_Master_Team_Yr_Wide, by = c("Team"))
Merge_Scholarship_Master_Team_Master_Wide1[is.na(Merge_Scholarship_Master_Team_Master_Wide1)] <- 0
Merge_Scholarship_Master_Team_Master_Wide2 <- merge(Scholarship_Master_Team_Pos_Wide2, Scholarship_Master_Team_Yr_Wide, by = c("Team"))
Merge_Scholarship_Master_Team_Master_Wide2[is.na(Merge_Scholarship_Master_Team_Master_Wide2)] <- 0
```
### Cluster Analysis
```{r scaling}
pr.out=prcomp(Merge_Scholarship_Master_Team_Master_Wide2[,-1], scale=TRUE)
```
## Clustering the Observations
```{r}
sd.data=scale(Merge_Scholarship_Master_Team_Master_Wide[,-1])
```



```{r hierarchical clustering}
par(mfrow=c(1,1))
~~~

~~~
data.dist=dist(sd.data)

plot(hclust(data.dist), labels=Merge_Scholarship_Master_Team_Master_Wide$Team,

main="Complete Linkage", xlab="", sub="",ylab="")

plot(hclust(data.dist, method="average"),

labels=Merge_Scholarship_Master_Team_Master_Wide$Team, main="Average Linkage",

xlab="", sub="",ylab="")

```

```{r cutting_clusters}

hc.out=hclust(dist(sd.data))

hc.clusters=cutree(hc.out,3)

table(hc.clusters,Merge_Scholarship_Master_Team_Master_Wide$Team)

```

```{r km_cluster}

km.out=kmeans(sd.data, 4, nstart=20)

km.clusters=km.out$cluster

table(km.clusters,hc.clusters)

```

```{r heirarchial_cluster}

hc.out=hclust(dist(pr.out$x[,1:5]))

plot(hc.out, labels=Merge_Scholarship_Master_Team_Master_Wide$Team, main="Hier. Clust.

on First Five Score Vectors")

table(cutree(hc.out,4), Merge_Scholarship_Master_Team_Master_Wide$Team)

```
~~~

~~~
```{r KMeans_Clusters}

# Employing kmeans clustering with different number of clusters and interpreting the results.

km.out2 <- kmeans(Merge_Scholarship_Master_Team_Master_Wide[,-1],2,nstart = 20)

km.out3 <- kmeans(Merge_Scholarship_Master_Team_Master_Wide[,-1],3,nstart = 20)

km.out4 <- kmeans(Merge_Scholarship_Master_Team_Master_Wide[,-1],4,nstart = 20)

km.out5 <- kmeans(Merge_Scholarship_Master_Team_Master_Wide[,-1],5,nstart = 20)

km.out6 <- kmeans(Merge_Scholarship_Master_Team_Master_Wide[,-1],6,nstart = 20)

```

```{r teamAssignment}

TeamNames <- c(Merge_Scholarship_Master_Team_Master_Wide[,1])

```

```{r km_result}

km.out5

```
~~~