Power 5 Clustering

Angela Ji, Matthew Bloom, Kenneth Marenco, Abhinav Ratnagiri

January 18, 2023

Contents

1 Problem Description

Many colleges and universities in the United States are grouped into one of the National Collegiate Athletics Association (NCAA) football conferences. Five of the Division I (D1) Football Bowl Subdivision (FBS) conferences are known as the Power 5: the Atlantic Coast Conference (ACC), the Big Ten Conference (BIG), the Big 12 Conference (Big 12), the Pacific Coast Conference (Pac-12), and the Southeastern Conference (SEC). There are also four additional D1 schools that are not members of a conference but are permitted to play in the FBS: United States Military Academy (Army), University of Notre Dame (Notre Dame), Brigham Young University (BYU) and University of Massachusetts, Amherst (UMass-Amherst).

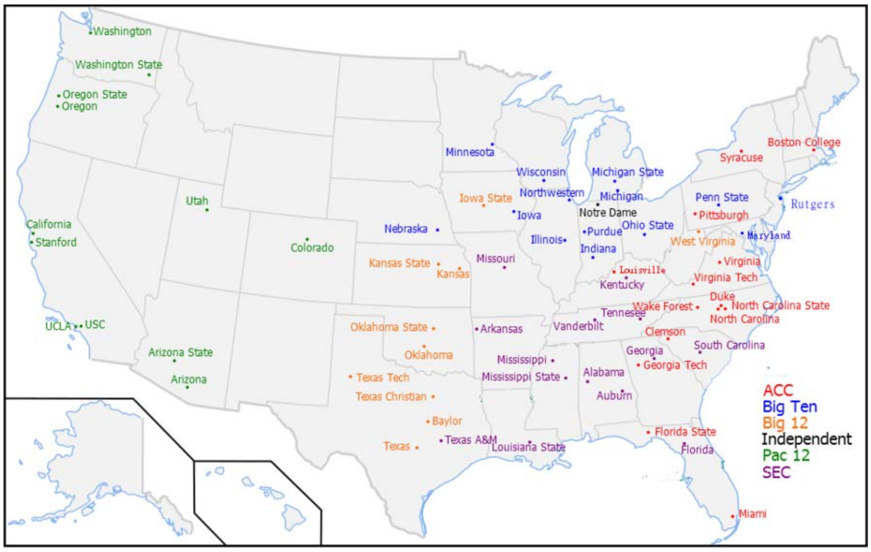

Figure 1: Map of the member schools of the actual Power 5 conferences.
Image from: https://en.wikipedia.org/wiki/Power_Five_conferences

The initial goal of this project is to utilize clustering methods learned in class to create five clusters of schools as similar to the original Power 5 as possible. In order to determine possible variables that were considered in the creation of the actual conferences, the team gathered a diverse set of data. The intent was to capture potential variables, both sports-related and not, that may be predictive of the Power 5 conference groupings. Ultimately, the goal of this process was to use this same method to categorize the independent schools. The same clustering algorithm was applied to the Power 5 schools as well as Army, Notre Dame, BYU, and UMass-Amherst, and they were all assigned a conference.

Next, a unique method is proposed that would cluster the schools into different groups than originally assigned in the Power 5, leading to five new conferences. The advantages and limitations of this grouping are discussed and compared to the actual school conferences.

This attempt at clustering the Power 5 conferences is an interesting problem because it highlights the disparities between data-driven clustering and actual groupings. Limitations of data-only clustering, such as the amount of publicly available data or obtaining outdated information, and the nuances of different clustering techniques, such as complete linkage versus single linkage, are revealed. Furthermore, non-data driven motivations, such as the pairing of rival schools together, can lead to clusters that are difficult or impossible to be represented using any data-driven clustering techniques. This divide between clustered conferences and actual school conference groupings can be observed in many areas and depicts how human decisions are never fully based on data alone. In practice, observations of the variables that unite these conferences might prove useful in the case of deciding how to assign schools that are new to the Power 5 (for example, Notre Dame recently became a voting member of the ACC).

2 Data Description

The five variables examined in attempting to predict the conference of each school are: hometown population, global position (latitude and longitude), federal graduation rate, historic number of National Football League (NFL) players, and student population. These categories were chosen because they included complete data for each school and describe diverse characteristics of each school, from education to professional sports alumni. The data is also indicative of qualities that the schools in each conference share, such as region, prestige, and strength of the football program. It was predicted that the geographic data would provide the strongest predictor of a school's conference (the names of several of the conferences are based on regions in the United States, after all). The associated summary statistics are presented in Table 1 below:

	Long.	Lat.	Num. NFL Players	Home Pop.	Grad. Rate 2011	Student Pop.
Mean	-92.019	37.678	266.324	350.074	69.284	32890.118
σ	14.199	4.546	104.619	697.790	10.248	13162.068
Min.	-123.279	25.718	36	6	47	4389
Max.	-71.169	47.751	599	3967	96	68679

Table 1: Summary statistics for the six variables that the group examined.

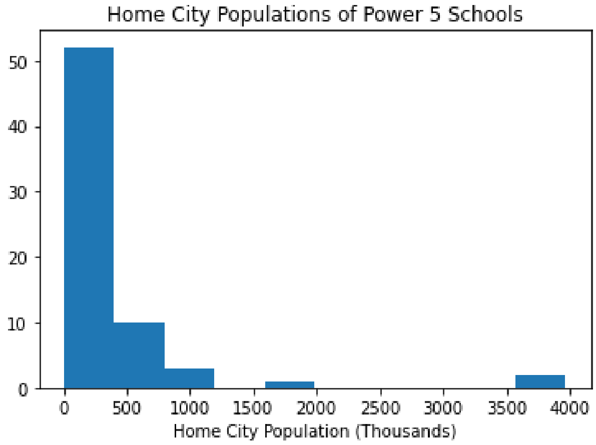

Figure 2: Distribution of Home City Population in the thousands. This histogram is skewed right with the majority of cities having populations less than 500,000 people and a few large cities with populations greater than 3,500,000.

Hometown population was collected by the United States Census Bureau which gathers voluntary data from citizens throughout the United States every ten years. The data corresponding to each of the schools in question was aggregated by the team.

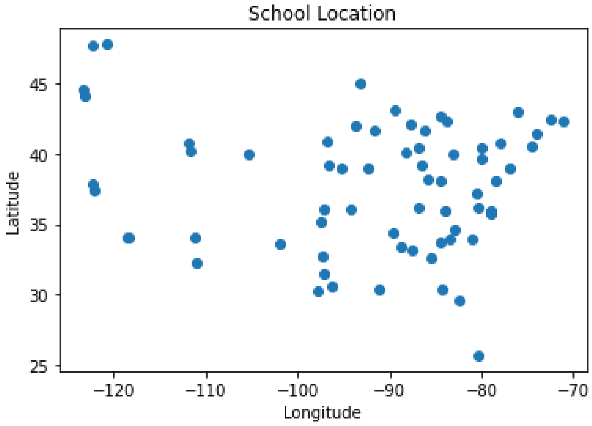

Figure 3: Scatterplot of school location by longitude and latitude.

The longitude and latitude of each school was aggregated from Google Maps data. While latitude and longitude is not accurate in measuring the physical distance between schools due to the Earth's curvature, it was sufficient in clustering using the difference between the degrees of longitude and latitude.

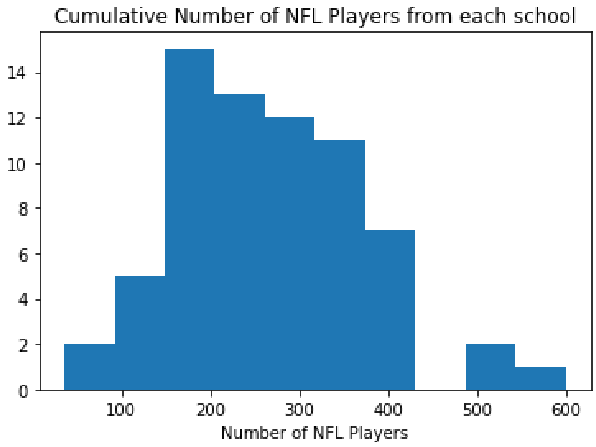

Figure 4: Distribution of number of NFL players from each school, historically.

The number of NFL players that have attended each school (through 2020) is aggregated from Pro Football Reference. This website uses data from ESPN Pro Football Encyclopedia which in turn tracks the institution each NFL player was officially enrolled in through publicly available university admissions data (if applicable).

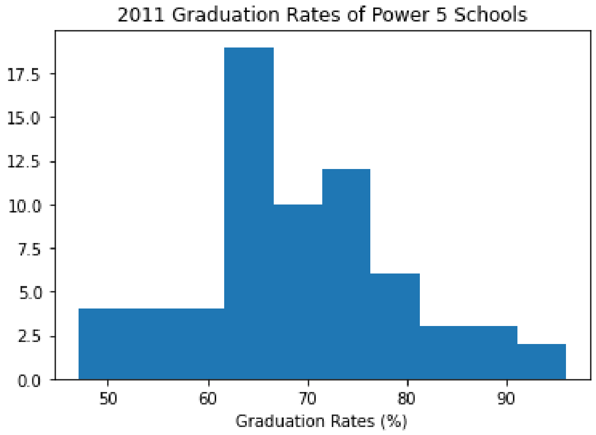

Figure 5: Distribution of the student graduation rate percentage in 2011.

The federal graduation rate in 2011 of each school's undergraduate population was taken from publicly available data from the NCAA. The NCAA maintains a record in which they collect yearly, mandatory, self-reported data from member institutions. The data was filtered for the most recent year collected (2011) as well as for the entire student body's graduation rate.

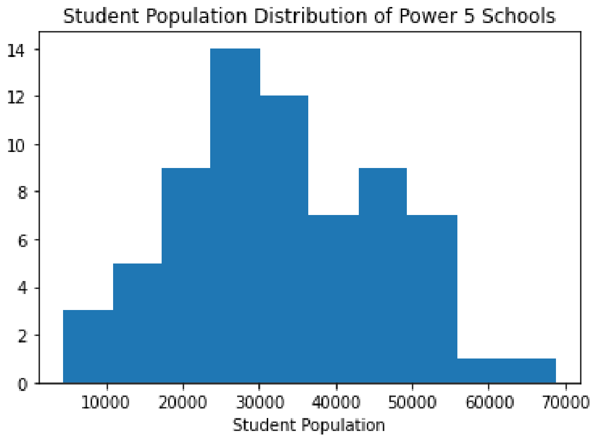

Figure 6: Distribution of Power 5 student populations in 2011.

The student population of each school was also taken from the NCAA's public data. Finally, the actual conference assigned to each school was also taken from data available from the NCAA. Fourteen schools compete in the ACC, ten in the Big 12, fourteen in the Big 10, twelve in the Pac 12, and fourteen in the SEC. There are four independent schools not affiliated with the NCAA but allowed to compete in the conferences.

3 Methods

All the possible combinations of the five variables (latitude and longitude were considered together, as "location", student population, hometown population, NFL players, and graduation rate) as predictive indicators of the conference were tested using four different clustering models: single, complete, and average linkage hierarchical clustering as well as k-means clustering.

Similarity between schools was assessed using the aforementioned combinations of the variables from the data collected. "Distance" between schools is defined by the N-dimensional Euclidean distance between the schools based on the variables' values, where N is the number of variables corresponding to each school that was used in a particular clustering (for example, in the clustering that considers hometown population and student population, N=2). All of our variables already had quantitative values, so none of them needed to be encoded as numbers.

Each clustering algorithm uses a unique distance metric to define its cost function. Ultimately, complete linkage hierarchical clustering was used, which attempts to minimize the maximal distance between clusters. In order to categorize the independent schools, we used the combination of variables that resulted in the highest predictive accuracy amongst all clustering models.

Since hierarchical clustering did not automatically label the clusters by conference name (the Python output only provides a number for each cluster, 0-4 in our case), labels were manually assigned based on the actual conference that had the most schools present in each cluster. For example, if the most common cluster label among the SEC schools was 2, all schools that were assigned to cluster 2 were assigned to the SEC. It was a mandatory condition that each conference was represented, as we needed an output that also had five conferences. This often meant that small clusters had to be labelled first to ensure uniqueness and then the other clusters would be labelled by majority school representation of the remaining unlabelled conferences. The predictive accuracy was then calculated as the total number of schools belonging to the actual Power 5 conference in each predicted conference cluster.

Since the variables were not separated by many orders of magnitude (the city populations were presented in units of thousands of people), it was not necessary to transform the data before running the clustering models.

4 Results

4.1 Part I

Complete linkage hierarchical clustering resulted in clusters that were most similar to the actual conferences while kmeans clustering produced clusters that were the most dissimilar. Multiple combinations of the five variables were analyzed for predictive accuracy, and latitude and longitude were found to yield the best results. This means that school location was the major factor in determining conference assignment with a few exceptions. Three examples of new conference clusters: Location only, Historic Number of NFL Players, and Student Population, can be found in the Appendix.

Conference	ACC Pred.	Big 12 Pred.	Big 10 Pred.	Pac-12 Pred.	SEC Pred.
ACC Actual	$\frac{9}{14} = 0.648$	0	0	0	5
Big 12 Actual	1	$\frac{6}{10} = 0.600$	3	0	0
Big 10 Actual	3	0	$\frac{10}{14} = 0.714$	0	1
Pac-12 Actual	0	4	0	$\frac{8}{12} = 0.750$	0
SEC Actual	1	2	1	0	$\frac{10}{14} = 0.714$

Table 2: Location Data Confusion Matrix

As seen in Table 2, location is a fairly accurate predictor of conference with a success rate of 43/64 colleges being sorted into their actual conference. The confusion matrix above shows that each conference achieved at least 60% percent accuracy with the Pac-12 being the best at 75%. This was the most accurate clustering result of the combinations attempted. Using multiple pieces of data such as the location and student population led to results that were less accurate than the location alone. Situations like this were the short falls of this data-driven task.

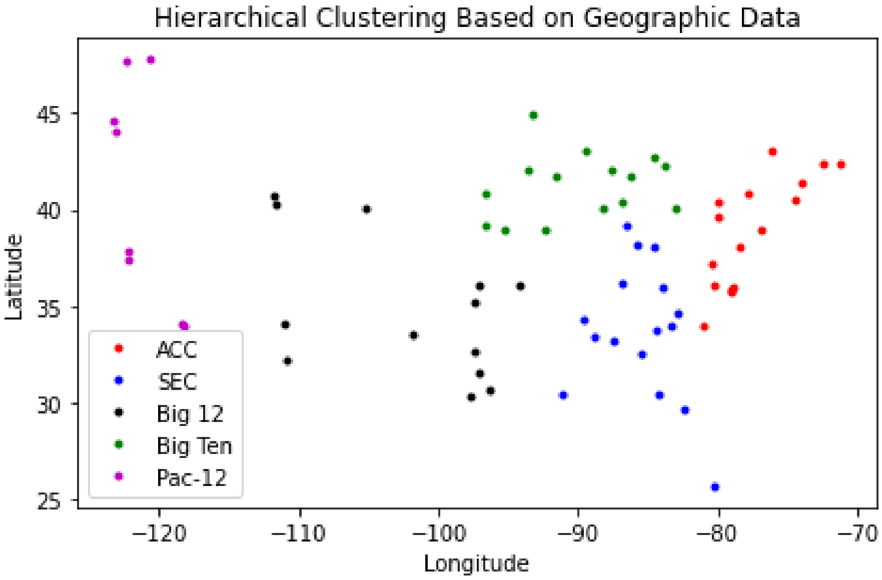

Figure 7: Hierarchical clustering based on geographic data

Figure 6 depicts the results of hierarchical clustering with complete linkage based on only the longitude and latitude of the schools. The scatterplot of this model is similar to the actual map of the conferences with the exception of a few outlier schools. Therefore, the Power 5 conferences are strongly related to location.

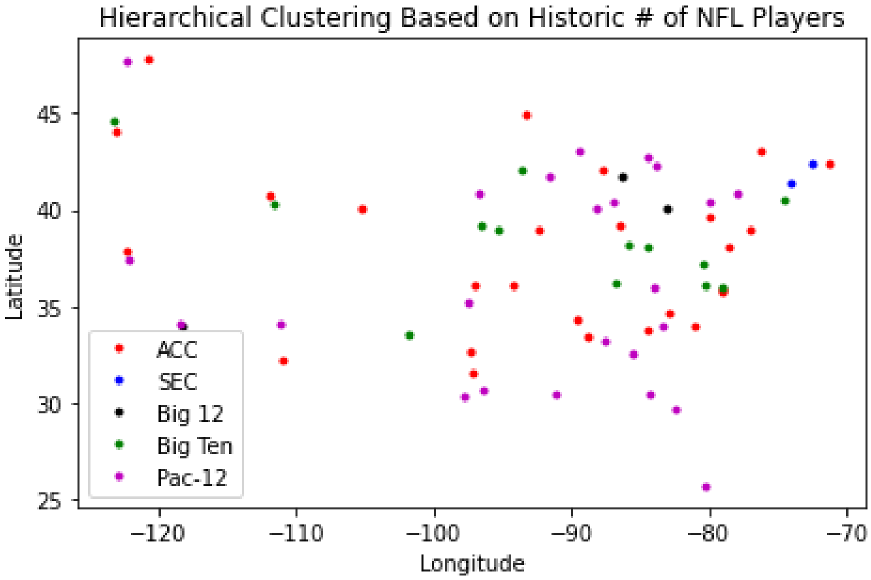

Figure 8: Hierarchical clustering based on historic number of NFL players

The next trials brought along lower accuracy than seen with the location data alone. Clustering based on the number of NFL players resulted in conferences that strayed further from the actual Power 5. In Figure 7 the groupings seem much more random than the actual conference clusters. One of the main issues seen is that there becomes an overwhelming number of schools that are placed into the ACC, while only two or three schools are assigned to the SEC and the Big 12.

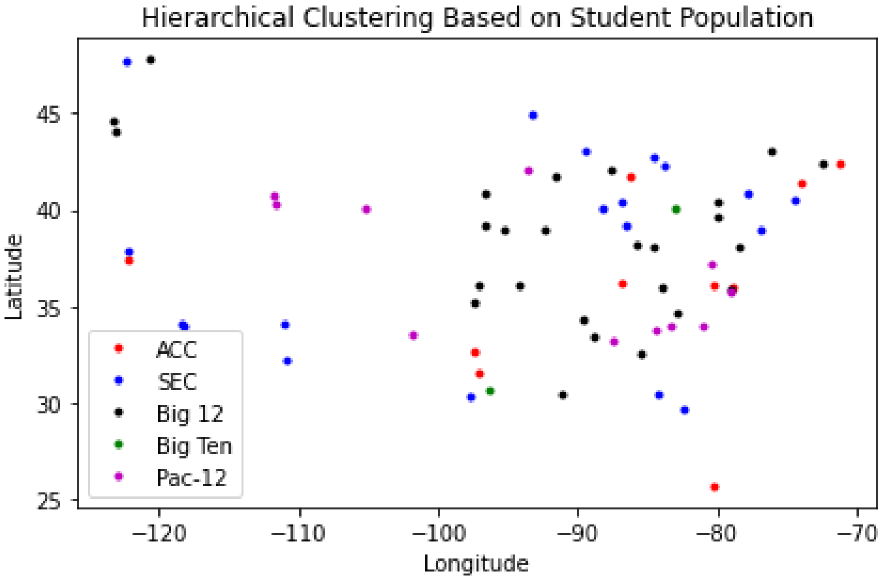

Figure 9: Hierarchical clustering based on student population

Clustering based on student population yielded similarly inaccurate results, as shown in Figure 8. The plot of predicted conferences results in seemingly random clustering when graphed by location data, and similarly uneven clusters emerge.

4.2 Part II

While testing different predictor variables in Part 1, we noticed that clustering by student population produced interesting results; although the clusters were not similar to the actual Power 5 conferences, they revealed an unique concept of matching schools by similar numbers of student populations. The schools competing in the new conferences may consequently have more similar stadium and audience sizes, making the training and game conditions more similar within conferences. Schools with similar size student populations may also have similar size athletic pools to choose from and spots open when scouting potential student athletes. Public school funding often partially depends on student population size, which would ultimately influence the athletic department funding and resources available to teams. The new conferences would then have more equal talent, funding, and play conditions, promoting more fair games. This new system is not without drawbacks however - one shortcoming that is clear from the map is the lack of geographic coherence within a conference, which would complicate travel for teams playing within their conferences. It also may lead to inequality in terms of conference attendance, with the larger schools in separate conferences from the smaller schools.

Clustering by location and student population size would lead to a large Pac-12 conference and a small SEC conference. This would generate more revenue for Pac-12 schools as well as increase the competition. It would be best for the SEC to be absorbed into a different conference, such as the Big 12, so that the two SEC schools could compete with more than just each other.

5 Conclusion

The distance metrics that were used provided a fairly accurate clustering system that was able to correctly analyze most of the colleges and their respective conferences. Location data was one of the best predictors, which makes intuitive sense as athletic teams within a conference must travel frequently in order to compete and thus conferences should be grouped to minimize travel time. During research on the formation of these conferences, one of the more difficult aspects to model was the date of formation. Schools shifted conferences frequently as other aspects of their situations changed, from revenue to athletic success. However, these shifts were often premeditated by a conference's need for change, and the school that was available at that time took the spot. This leads to disparities between purely location-based clusters and the actual conferences, as seen in the example of Rutgers University being in the Big Ten whereas the University of Pittsburgh is in the ACC. These schools geographically are very close, but based on other factors, they were assigned to different conferences.

Something that worked well in creating mostly accurate clusters was the use of complete hierarchical linkage. k-means wasn't a good clustering method for this dataset because many of the schools across conferences had similar metrics, but schools within the same conference differed greatly. This resulted in one or two very large clusters while a few outlier schools made up the other clusters. Therefore, k-means led to an algorithm failure, as it performs best when clusters are distinct and concentrated. Complete linkage was the best of the hierarchical clustering methods because it created clusters of similar size based on the smallest distance between the farthest points of the clusters. Since single linkage minimizes distance from the closest point of each cluster, there would often be one cluster that contained a majority of the schools.

If this analysis were to be repeated, a more complex method of clustering could be employed that assigned different weights to different variables. This would allow latitude and longitude to be heavily weighted since they are the most predictive variables of the conference while other variables may only change the predicted conference of edge cases. More data would also increase the number of possible combinations of variables used to predict the conference, and variables that were not included in the initial dataset may ultimately result in more accurate clusters. In particular, it may be interesting to explore economic variables more directly in the future, as the business side of sports intuitively seems to be a strong factor in grouping schools. Variable weighting combined with more data would enable more refined manipulation of the predicted conference in order to more closely match the actual Power 5.

When analyzing models with variables other than location included, oftentimes conferences would be generally clustered together relative to their actual conference, but two conferences may fall under one predictive cluster, making it larger than average, while another cluster would have only a few outlier schools. If these variables were to be used to create unique conferences as outlined in Part II of the problem statement, they would distribute schools more evenly between clusters if the total number of clusters was variable. A small cluster with only a few schools would be illogical because there would not be much competition, so merging the small cluster with another cluster would reduce the disparity between the number of schools in each conference.

References

[1] Bureau, US Census. City and Town Population Totals: 2010-2019. The United States Census Bureau, 21 May 2020, www.census.gov/data/datasets/time-series/demo/popest/2010s-total-cities-and-towns.htmltables.

[2] College Football Encyclopedias and NFL Records. Pro Football Reference, www.pro-football-reference.com/schools/.

[3] Pbrock. “Football Records Books (since 2004).” NCAA.org - The Official Site of the NCAA, 27 Mar. 2020, www.ncaa.org/championships/statistics/football-records-books-2004.

[4] NCAA.com. “NCAA College Football FBS Stats.” NCAA.com – The Official Website of NCAA Championships, NCAA.com, 4 Apr. 2021, www.ncaa.com/stats/football/fbs.

[5] Rpowell. “Shared NCAA Research Data.” NCAA.org - The Official Site of the NCAA, 24 Sept. 2020, www.ncaa.org/about/resources/research/shared-ncaa-research-data.

[6] ”Sklearn.cluster.AgglomerativeClustering.” Scikit, scikit-learn.org

[7] “Power Five Conferences.” Wikipedia, Wikimedia Foundation, 4 Apr. 2021, en.wikipedia.org/wiki/Power_Five_conferences#Current_conferences_and_teams.

A Conferences Clustered by Location (Latitude and Longitude)

ACC	**Big 12**	**Big 10**	**Pac-12**	**SEC**
Virginia	Oklahoma	Kansas	Washington	Georgia Tech
NC State	UT-Austin	Iowa State	S. California	Florida State
Virginia Tech	Texas Tech	Kansas State	UCLA	Clemson
Pittsburgh	Texas Christian	Purdue	Washington State	Louisville
Syracuse	Baylor	Minnesota	UC Berkeley	Miami
Chapel Hill	Oklahoma State	Iowa	Oregon State	Indiana
Boston College	**Brigham Young**	Illinois	Stanford	Auburn
Wake Forest	Utah	Nebraska		Louisiana State
Duke	Arizona State	Michigan		Vanderbuilt
West Virginia	Colorado	Michigan State		Mississippi State
UMD	Arizona	Ohio State		Mississippi
Penn State	Texas A&M	Wisconsin		Kentucky
Rutgers	Arkansas	Northwestern		Florida
Army		**Notre Dame**		Georgia
UMass-Amherst		Missouri		Alabama
S. Carolina				Tennessee

Table 3: Conference member schools based on location (independents in bold).

Accuracy: ACC: 9/14, B12: 6/10, B10: 10/14, P12: 8/12, SEC: 10/14

Total: 43/64 67.2%

B Conferences Clustered by Historic Number of NFL Players

ACC	**Big 12**	**Big 10**	**Pac-12**	**SEC**
Virginia	Texas Christian	Syracuse	NC State	S. Carolina
Pittsburgh	**Notre Dame**	Chapel Hill	Virginia T	Rutgers
Georgia Tech	Florida	Texas Tech	Florida State	
Clemson		Kansas State	Louisville	
Boston		Maryland	Miami	
Wake Forest		Illinois	Kansas	
Duke		Wisconsin	Oklahoma	
West Virginia		Northwestern	UT Austin	
Iowa State		**UMass-Amherst**	Baylor	
Minnesota		Washington	Oklahoma State	
Iowa		Arizona State	Purdue	
Nebraska		Colorado	Indiana	
Michigan		Missouri	Penn State	
Michigan State			S. California	
Ohio State			Washington State	
Army			UC Berkeley	
Brigham Young			Oregon State	
Utah			Stanford	
Oregon			LSU	
Arizona			Vanderbilt	
Auburn			Kentucky	
Miss. State			Arkansas	
Texas A&M			Alabama	
Mississippi			Tennessee	
Georgia				
UCLA				

Table 4: Conference member schools based on number of NFL players through 2020 (independents in bold).

Accuracy: ACC: 5/14, B12: 1/10, B10: 4/14, P12: 5/12, SEC: 1/14

Total: $16/64 = 25\%$

C Conferences Clustered by Student Population

ACC	**Big 12**	**Big 10**	**Pac-12**	**SEC**
NC State	Boston	Florida State	Virginia	Ohio State
Virginia Tech	Wake Forest	UT Austin	Pittsburgh	Texas AM
Georgia Tech	Duke	Purdue	Syracuse	
Iowa State	Miami	UMD	Chapel Hill	
Texas Tech	Texas Christian	Minnesota	Clemson	
Brigham Young	Baylor	Illinois	Louisville	
Utah	**Notre Dame**	Indiana	Kansas	
Colorado	**Army**	Penn State	Oklahoma	
S. Carolina	Stanford	Michigan	West Virginia	
Georgia	Vanderbilt	Michigan State	Oklahoma State	
Alabama		Wisconsin	Kansas State	
		Rutgers	Iowa	
		Washington	Nebraska	
		S. California	Northwestern	
		Arizona State	**UMass-Amherst**	
		UCLA	Oregon	
		UC Berkeley	Washington St.	
		Arizona	Oregon St.	
		Florida	Auburn	
			LSU	
			Miss. St	
			Mississippi	
			Kentucky	
			Arkansas	
			Missouri	
			Tennessee	

Table 5: New conferences based on student population (independents in bold).

Accuracy: ACC: 3/14, B12: 2/10, B10: 10/14, P12: 3/12, SEC: 1/14

Total: 19/64 = 29.7%