Predicting College Football Success from High School Statistics

by Eric Arnett, Ritwik Roy, Michael Ferrin, Connor Gullstad

Introduction:

When looking over statistics for high school football players, our group noticed that many of the players that achieved high levels of success and national prestige at the high school level did not always perform as well as expected in college, and conversely, that some low ranked high school players performed very well at the college level. After taking this observation into consideration, our team would like to compare statistics of nationally ranked high school football players, and compare it to their college performance statistics, to see how strongly the high school rankings predict a player's college performance.

First, we will extract HTML code from rivals.com on the player rating of ranked high school players from years past, and their corresponding college offensive statistics from ESPN.com, to create a database of raw data which we will use to analyze. Then after cleaning and organizing the data into a form that we can easily manipulate, we will examine the college performance of different offensive skill players - Quarterbacks, Running Backs, and Receivers - based off of their total offensive yardage, and total touchdowns, and compare it to their high school Star rating (a scale from 1 to 5, 5 being the best) to assess if they perform as well in college as expected. We will also look at regional differences to show if high school players in different parts of the U.S. are producing better performing college athletes.

Data Extraction and Cleanup:

We extracted our data from primarily two sources: Rivals.com and ESPN.com. Our high school player data was from the graduating class of 2010, and our college data was from the 2012 college football season. We chose these years because they are fairly recent and thus may more accurately predict the correlation between player rating and college performance in a contemporary setting than from farther in the past. We also chose to gather college data from 2012, their third season, rather than the senior year season data from 2013 because several players leave to play in the National Football League after their junior season (the first year they are eligible to do so), and thus there is no senior year data for many of the players.

Rivals.com provided our high school player data, and ESPN.com provided the corresponding college statistics. From both websites, we located the source code of online HTML data we wished to select, saved the data, and viewed it in Rstudio using the readhtmltable() function from the XML package. We then merged high school and college data tables based on player name, and saved our final cleaned dataset as a CSV file with the data on Name, Position, Home State, High School Star Rating, and total yards and touchdowns from the 2012 season:

With this properly cleaned data set, we ran a number of visual and analytical tests in R.

Data Analysis:

With the wealth of data at our disposal, we calculated a large variety metrics with respect to position, player rating, and region. Each player is classified by their attributes in regards to the three aforementioned descriptors. Players could play position Quarterback (QB), Running Back (RB), or Wide Receiver (WR); players could have a rating of 2 stars, 3 stars, 4 stars, or 5 stars; and, players were classified by the location of their high school as part of four distinct regions in the United States: West (Alaska, Hawaii, Washington, Oregon, California, Idaho, Nevada, Utah, Arizona, Montana, New Mexico, Colorado, and Wyoming), Midwest (North Dakota, South Dakota, Nebraska, Kansas, Oklahoma, Texas, Minnesota, Wisconsin, Iowa, Michigan, Illinoi#s, Indiana, Ohio, and Missouri), South (Kentucky, Tennessee, Arkansas, Louisiana, Alabama, Mississippi, Georgia, Florida, South Carolina, North Carolina, Virginia, and West Virginia), and Northeast (Pennsylvania, Maine, Vermont, New Hampshire, Massachusetts, Connecticut, Rhode Island, New York, Maryland, Delaware, and New Jersey). We analyzed players according to their positions, ratings, or geographical origins in isolation in addition with respect to the intersection of two of these classifications.

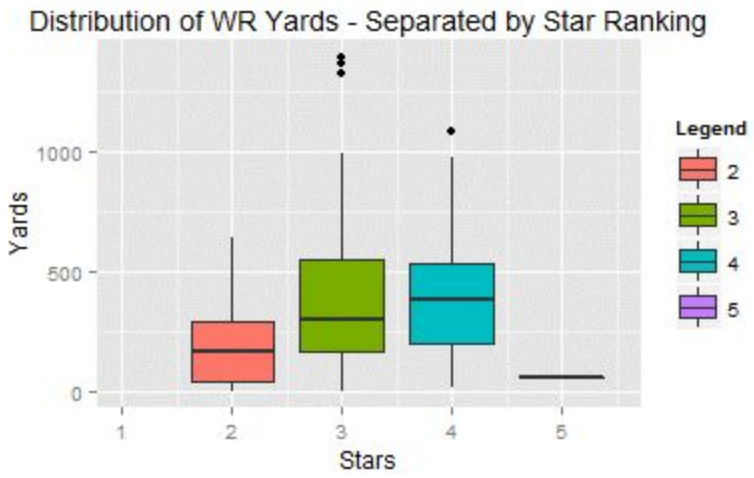
The analysis centered around looking at statistical metrics of players by yards and touchdowns (TDs). Calculating averages, using the sum() and nrow() functions, for these statistics provides a benchmark level of expected performance. These averages were calculated for individual positions, as the expected number of yards thrown for by a quarterback is expected to be different than the number of yards rushed for by a running back or the number of receiving yards by a wide receiver. After attaining a performance benchmark through calculating average

yards and touchdowns for each position, the logical next step was to calculate the standard deviations, using the sd() function, of each of these six statistics (yards and touchdowns for each of the three positions). Knowing standard deviations provides an idea of the distribution of performance. Then, we were able to calculate, for each player, how they performed in yards and touchdowns in terms of standard deviations above or below the mean. This step was critical, because it allowed us to compare players across different positions groups through a normalized metric (performance in standard deviations above or below the mean), independent of statistical expectations for their position. Once we had this ability, we began analysis of three double-variable relationships: player ratings vs. region, player ratings vs. position, and region vs. position.

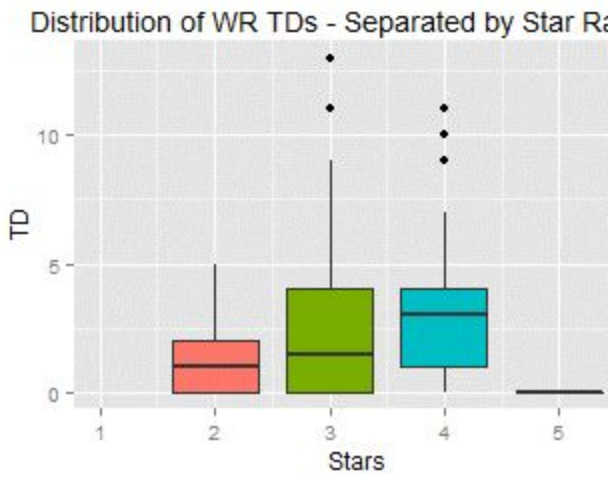
Distribution of TD's and Yards By Position

The main metrics used to assess offensive performance in football are the amount of total yards, and amount of total touchdowns in a season. The following box and whisker plots displays the minimum, lower quartile, average, upper quartile, and maximum number of touchdowns thrown by players based on star rating and position:

Wide Receiver Distributions



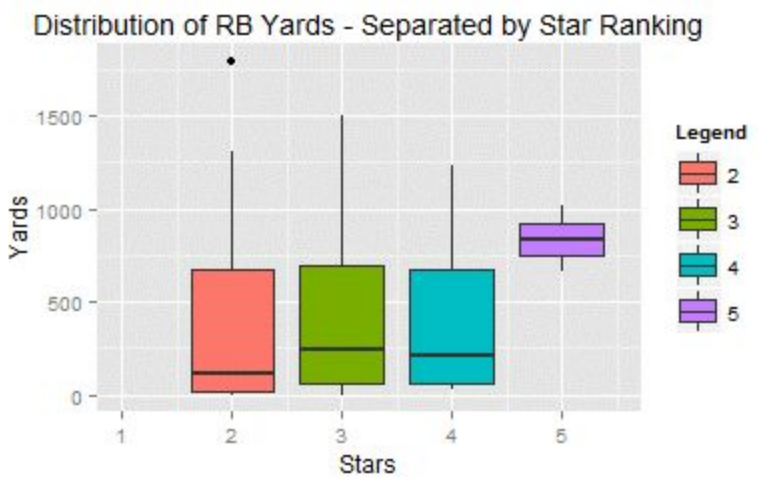
Based on the above graph, it appears that there is definitely some sort of correlation between total offensive yards in a college season and star rating for Wide receivers. It is clear that 2 Star quarterbacks have a much lower overall distribution of yards at all quartiles than 3 or 4 star receivers. However, the difference between 3 and 4 star receivers is much smaller. Although 4 star receivers had a higher median amount of total yards that 3 stars, the difference is very small, and the 3 star receivers have many outliers with over 1,000 total yards, much more than the four star receivers. Overall, this data seems to suggest that the both 3 and 4 star receivers throw have a much higher total yardage than 2 stars on average, but there is little difference between the yardage that will be thrown by 3 or 4 star recruits in colllege. There was only one 5 star receiver in the class of 2010, so the data is inconclusive on how 5 stars perform.



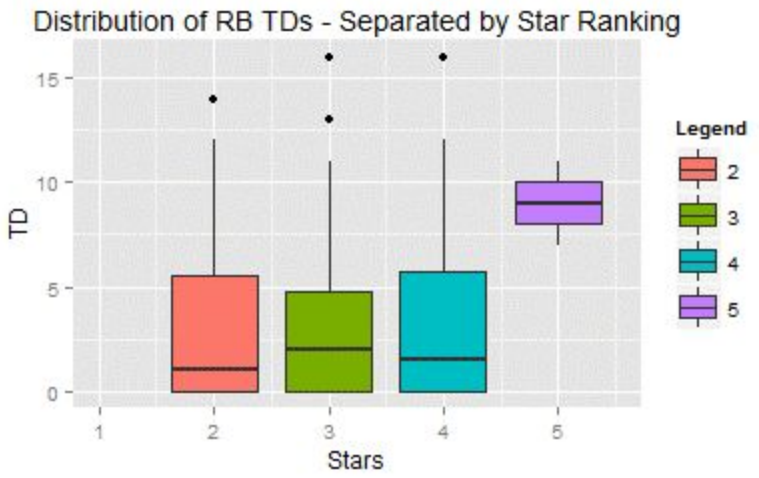
Much like the distribution of yardage for wide receivers, it appears that there is a strong difference between the amount of touchdowns scored in a college season by 2 star high school recruits players versus 3 and 4 star recruits, but that the difference in TD's scored is much smaller and harder to predict between 3 and 4 star recruits. 4 star recruits have a higher median TD's scored, and an overall lower range, suggesting their performance is slightly more predictable than 3 stars. However at the same time, the two highest scoring players were both 3 star recruits, as shown by the large range and upper outliers.

Between the two graphs, there overall seems to be some correlation between high school star rating and offensive performance for Wide Receivers. 4 star recruits appear on average better than 3 star recruits based off of the median, but 3 star recruits have a higher number of outliers that are very high performers than 4 star recruits. 2 star players are predictably lower in all categories

Running Back Distributions

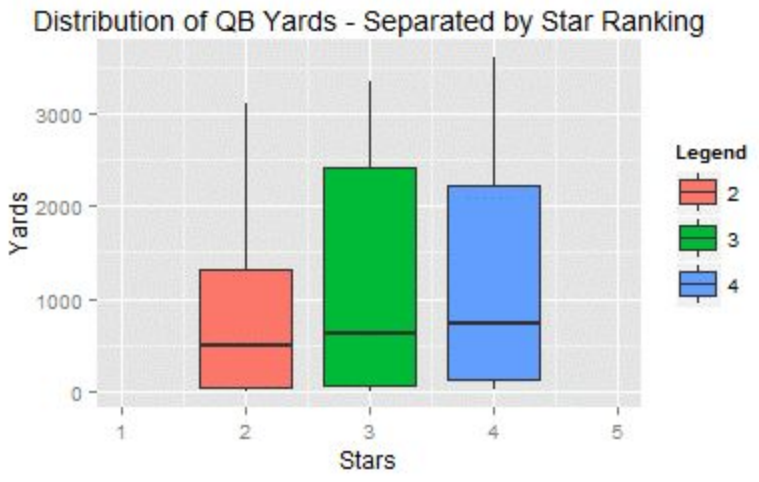


The difference in yardage based off of star ranking for Running Backs is much more subtle for all star ranks than for Wide Receivers. 2 star running backs do have the lowest quartiles, and maximums, but by a very, very slight amount. Additionally the player with the highest overall yardage in the 2012 season was a 2 star recruit, as shown by the outlier (This recruit was Le'Veon Bell, now the starting running back for the Pittsburgh Steelers). 3 star recruits actually performed slightly better than 4 star recruits in terms of median and max, but the difference is subtle. 5 star recruits had the highest overall average by far, but like wide receivers there is not enough data to make this a predictable marker.

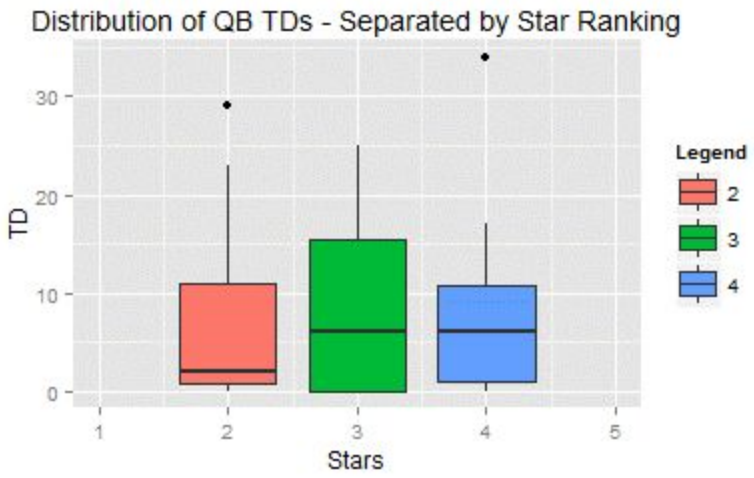


Like the distribution of yardage, the distribution of touchdowns shows very little difference between 2, 3 and 4 star recruits. Suggesting that for Running Backs, there seems to be little correlation between the high school star rating and success in college.

Quarterback Distributions



Quarterback's are similar to running backs in that there is relatively little difference between the high school star rating and total season yardage. The median yardage does increase slightly with each star rating though, and the upper quartile for 2 star recruits lags far behind 3 and 4 star recruits. There were no 5 star quarterbacks in the class of 2010.



Touchdowns thrown tell a similar story as yardage does - that it does not appear that star rating has much of an effect on offensive productivity for Quarterback's. 2 star recruits did, however have a much lower median and range than 3 or 4 star recruits, a the 4 star outlier was the highest overall performer (Tyler Bray), and three star recruits had a higher upper quartile.

---Overall Analysis---

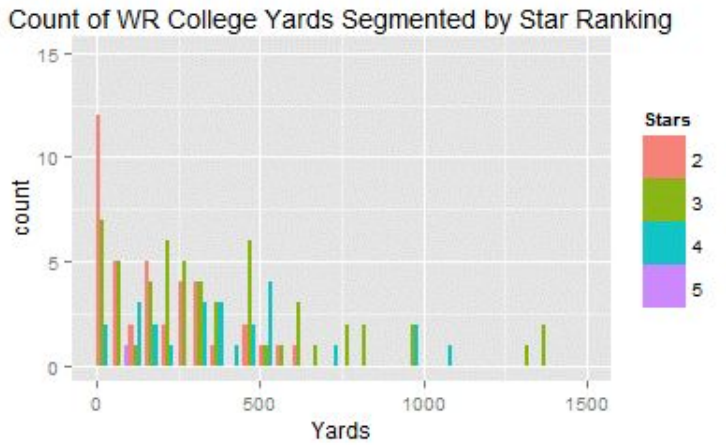
These distributions suggest that star rating is only a decent predictor for college performance for wide receivers, and not for running backs or quarterbacks. Specifically, 2 star wide receivers had much lower amounts of yards/TD's than 3 or 4 star receivers, but that there

was still little difference in performance between 3 and 4 star receivers, so overall star rating was not a very clear indicator of offensive performance based off of these box plots.

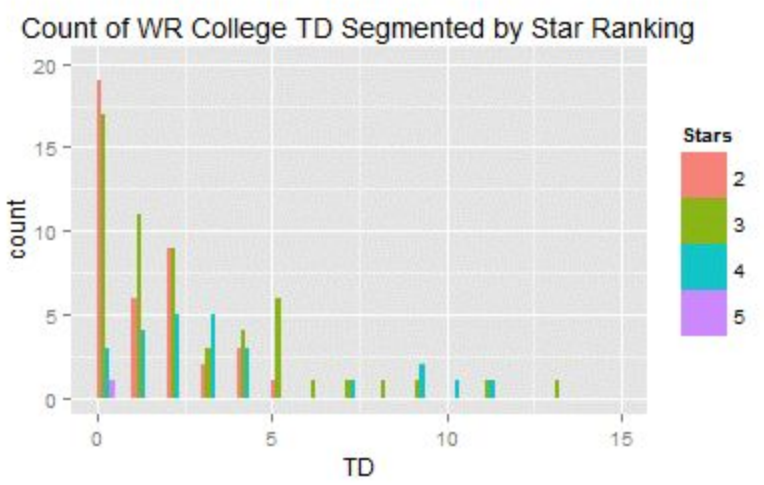
Count of Total Yards/TD's by High School Star Rating

The following segmented histograms display the number of players who were within a certain range of TD's scored or Total Yards (the bins represent the range). The different colored segments of the bars represent the different high school star ratings. One thing to note with regard to these graphs is that the number of 2, 3, 4 and 5 star receivers is not evenly distributed amongst the whole player pool. In general, because it takes more skill and work to receive a higher high school player rating, we would expect there to be less players with a given star rating as the star rating increases. This is evident in the fact that there are almost no 5 star recruits (these players are the rarest and best talents in the country). This may alter the way the data looks because the numbers are seen as absolute counts, and not proportionate counts.

Wide Receiver Counts



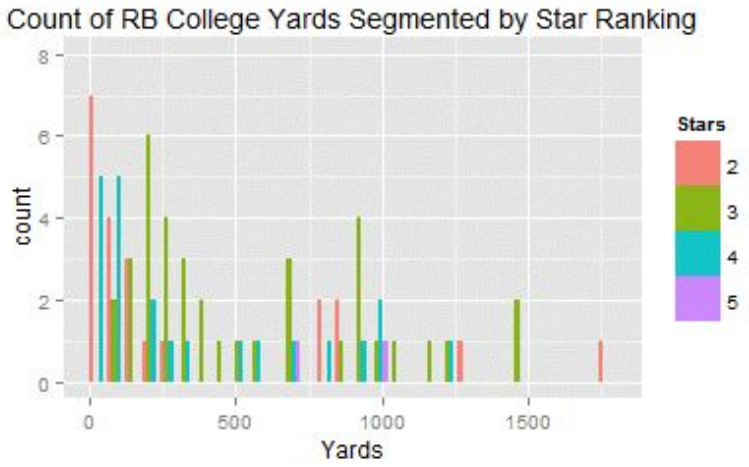
The above graph shows how many of 2,3,4 and 5 star receivers scored within the bin range of yards. The data seems to suggest a trend that as the yardage increases, so do the proportion of players with higher star ratings that had yardage within that bin - the overall amount of players also tends to decreases as the yardage increases, as shown by the rightward skewed graph. The lowest yardage bin seems to be the most similar to what we would expect, with 2 star receivers having a much higher presence than 3 star receivers, which in turn have a higher presence that 4 stars. 2 star receivers were highly represented in the lower yardage bins (<250 yards), however as the yardage increased, 3 and 4 star receivers became more highly represented.



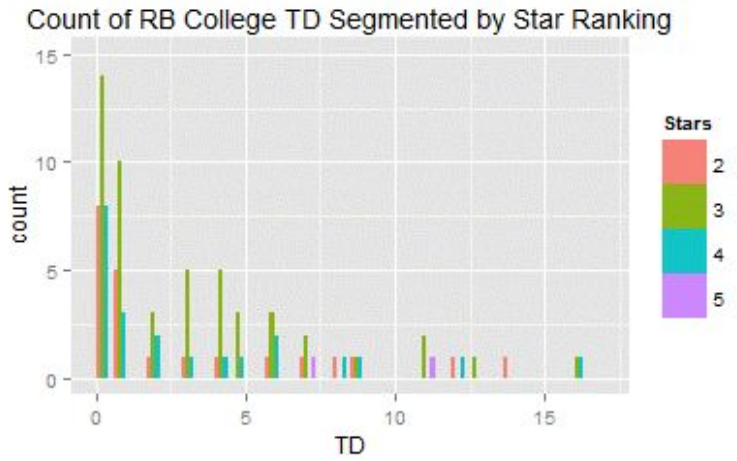
The count for TD's shows a similar trend as the yardage count. In this case however, 2 star WR's struggled to even score one touchdown in a given season, as shown because 3 star receivers dominated the proportion of players with 0-1 TD's. In general though, as yardage increases so does the average star rating within each bin. It appears that in most bins, even the

ones farthest to the right, that 3 star receivers are more highly represented than would be expected, and often represent a higher proportion of the total player count than 4 star receivers, based on star count.

Running Back Counts

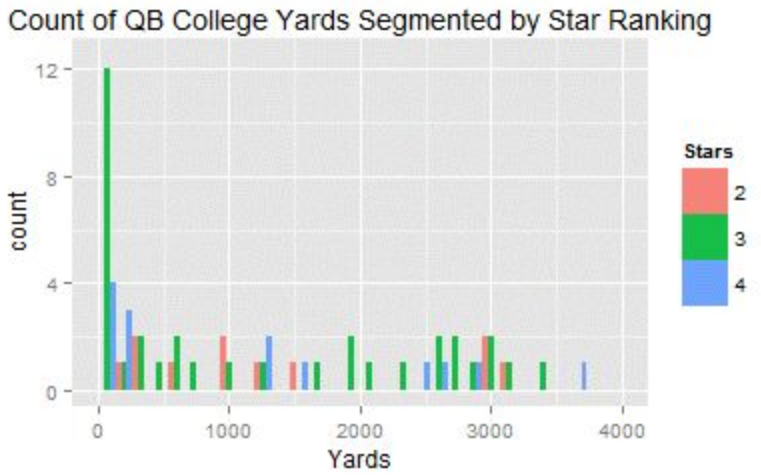


Running backs show a similar trend that receivers do in terms of yardage, but the data is slightly less linear. In general, as the yardage bins increase, 3 and 4 star backs become more proportionately represented. However unlike with Wide Receivers, 2 star running backs have a much larger distribution of yardage throughout the graph - with the 2 star high school back Le'Veon Bell showing up as the best college RB in terms of yards.

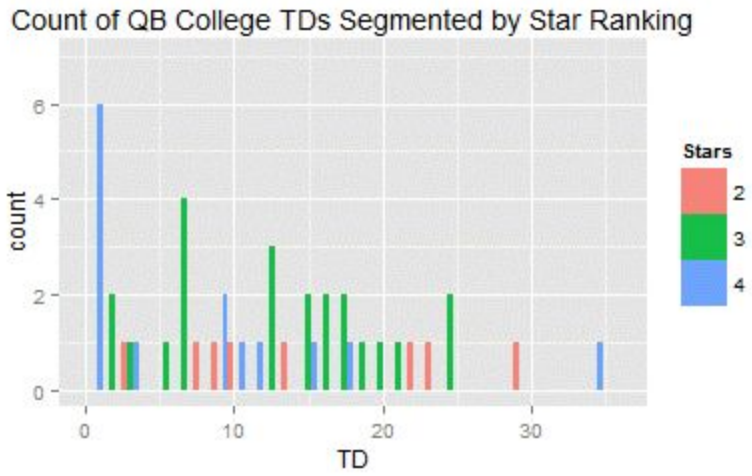


The data for TD counts appears rather inconclusive for Running Backs. In general, it appears that 3 star receivers (green) are very prevalent amongst all different TD counts. This is may be because there are a both a higher number of 3 star RB's in the NFL than 4 star players, but there performance is almost on par with that of the 4 star backs. Thus, they are heavily represented in all areas on this graph.

Quarterback Counts



The most striking element of the above graph of yardage segmented by star ranking for Quarterbacks is the extremely high count of 3 star QBs in the lowest bin level, and that the clear second highest count was 4 star QBs, although it is much less prevalent than the 3 star players. This is contradictory to what we would expect if we assume that star rating corresponds to higher performance - that 2 stars would be the most prevelent at this bin level. The rest of the data is fairly inconclusive.



This last graph for of the quarterback touchdown count is even stranger, with 4 star players having by far the highest count of players with very few touchdowns. This is especially strange because 4 star players are the least prevalent in terms of total count (except for 5 star players). 3 star players seem to have high counts in several different touchdown counts.

Overall, these graphs of "counts" does not provide hugely strong evidence that high school star rating correlates to college offensive productivity. Much like the boxplot distribution

graphs, the most convincing evidence was with Wide Receivers. For both yardage and touchdowns, as the as the numbers increased, so did proportion of more highly ranked players, on a fairly consistent level.

Regional Recruiting Talent

A hot topic in debate of high school and college football is which region of the country produces talent of the highest quantity and quality. For instance, many sportswriters attribute the historical and recent success of the Southeastern Conference (SEC) to its geographical positioning in the southeastern portion of the country, which is believed to be a recruiting hotbed. To investigate these claims and ascertain other information, we ran the numbers.

The first figure of note is simply the raw number of recruits from each region (West, Midwest, South, or Northeast—as defined previously) regardless of player rating or position. There were 60 players from the West, 89 from the Midwest, 121 from the South, and 17 from the Northeast. We didn't feel the need to control for the relative populations of these regions, as the states these regions are comprised of are generally the recruiting areas for schools within those regions. The immediate takeaway here is that the South does indeed produce the most talent in absolute numbers. The West and Midwest regions produce a moderate amount of talent, while the Northeast lags quite far behind. Looking at the distribution of 2-, 3-, and 4-star players by region (only three 5-stars provided an insufficient sample size) showed that the actual talent level of players (as indicated by player rating) was distributed through regions in similar proportions to the overall regional distribution, indicating that regions differed not in being top- or

bottom-heavy in player talent, but rather, simply, in the absolute amount of player talent produced. This allotment of players by region affirms the beliefs of many that the South produces the largest amount of football talent, and likely provides a reason for the recent dominance of SEC teams in college football.

Conclusion:

If I were a college football coach, scouting out talented players, I would try to focus less on how well the player is ranked and more so on potential, and on how well they fit with the team. Star rating does seem to be a decent indication of how the player will perform at the next level, but it is by no means the end all be all and should not be the deciding factor when selecting a player. The one notable exception to this is for Wide Receivers. There seems to be a decent correlation that 3,4, and 5 star wide receivers perform much better in college than 2 star receivers. 2 Star receivers throw far lower TD's and have less offensive yards than more highly ranked players. Moreso, players in the Southeastern Region consistently high a higher player rating which seems to translate well into college play, so scouting for players from these regions is advisable if possible.

This project allowed our team to put our skills with data collection, cleaning, and analysis with R to the test. Hopefully our work will be useful for the Golden Bear Football team when selecting its recruits. You can thank us when we win the Rose Bowl!

Sources:

http://sports.yahoo.com/ncaa/football/recruiting/recruit-search
http://espn.go.com/college-football/statistics/_/year/2010