# Is The Playing Field Really Level?

## An Analysis Comparing The Rankings Of Top High School Football Players To The U.S. Distribution Of Income

**Authors: Greg Holden, Jacob Schmidt and Tawfiq Zureiq**

## Background

Ranking the top U.S. high school football players has become a big business. Colleges rely heavily on these rankings to allocate athletic scholarships. Receiving a scholarship to a top college program often creates a pathway to a professional career in the NFL. NFL contracts – especially for 'premium' positions such as quarterback – are extremely lucrative.

An average of ~3,300 incoming college players are ranked each year. Ratings are assigned based on a scale of 1 (lowest) to 5 (highest) stars. 4-star prospects represent the top ~10% of all prospects, while only ~1% of all prospects receive a 5-star rating.

But are these star rankings equally distributed? Or do wealthier players with access to private coaching and specialized camps while in high school (or earlier) tend to receive a disproportionate share of the top star ratings?

### Average Players Per Star Ranking: 2011-23

| | Stars | | | | | |
|---|---|---|---|---|---|---|
| | **1** | **2** | **3** | **4** | **5** | **Total** |
| # of Players | 74 | 1,115 | 1,775 | 328 | 32 | 3,324 |
| % of Total | 2.2% | 33.5% | 53.4% | 9.9% | 1.0% | 100.0% |

## Motivation

We want to investigate whether high school recruits across the income spectrum tend to receive proportionate shares of top (5 and 4) star rankings.

Previous literature has focused on the relationship between star rankings and team success (Mankin, Rivas & Jewell, no date), the monetary value each star ranking provides to a college (Bergman & Logan, 2020), and how star rankings relate to NFL success (Wheeler, 2018).

However, we are unaware of any research regarding the link between average income and assigned star rankings. Our goal is to fill that gap.

## Project Objectives

We have several specific research topics for this project:

- Analyze how the star rankings of top high school players compare to income levels across the U.S.
- Establish whether this effect is more pronounced for specific positional groups.
- Analyze whether certain colleges tend to recruit more heavily from certain income segments

# Is The Playing Field Really Level?

## Description Of Data Sources

**Our analysis used three different data sources – one to provide data on each college football player, a second to provide zip code information, and a third to provide average income data.**

### CollegeFootballData.com

- API is located at: https://api.collegefootballdata.com/api/docs/?url=/api-docs.json
- Querying API returns a JSON object
- **Primary variables**:
    - **Name** | Recruit's name
    - **Stars** | 1 - 5 Recruiting rating signifying the quality of a recruit, with 5 being the best
    - **CommittedTo** | Recruit's choice of college
    - **Position** | The spot the recruit plays on the team
    - **Height** | Height of recruit
    - **Weight** | Weight of recruit
    - **City** | City the recruit is from
    - **StateProvince** | State recruit is from
    - **HometownInfo.longitude** | Longitude of recruit's hometown
    - **HometownInfo.latitude** | Latitude of recruit's hometown
- 29,435 records returned
- Years used: 2011 - 2023

### Zip Code Database

- Free CSV download is located at https://www.unitedstateszipcodes.org/zip-code-database/
- **Primary variables**:
    - **ZIP** | ZIP Code used for merging with income information
    - **Primary City** | City associated with ZIP Code
    - **State** | State associated with ZIP Code
    - **Latitude** | Latitude associated with ZIP Code
    - **Longitude** | Longitude associated with ZIP Code
- 42,735 records
- Time period is not relevant

### Individual Income Statistics

- Free CSV download is located at https://www.irs.gov/statistics/soi-tax-stats-individual-income-tax-statistics-2019-zip-code-data-soi
- **Primary variables**:
    - **State** | State of income information
    - **ZipCode** | ZIP Code of income information
    - **A02650** | Total income amount
- 1,011,577 records
- Years used: 2015 - 2020

# Is The Playing Field Really Level?

## Key Data Manipulation Steps

### Overview

Given that our end goal was to analyze the relationship between income and college football star rankings, we needed to be able to merge the income data with the recruiting data. Given our available variables, we chose the city where the athlete is from to merge our income and recruiting data. As you can see from the previous slide, the i**ncome dataset has no column for city.** Luckily, our ZIP Code database has this necessary information.

### We first needed to merge the Income and the ZIP Code databases using ZIP.

**Zip Code Database**

| state | zip | primary_city | latitude | longitude |
|---|---|---|---|---|
| MN | 55101 | Saint Paul | 44.95 | -93.09 |

**Income Database**

| STATE | zipcode | A02650 |
|---|---|---|
| MN | 55101 | 13627.0 |
| MN | 55101 | 40567.0 |
| MN | 55101 | 55233.0 |

**Combined Database**

| state | zip | primary_city | latitude | longitude | STATE | zipcode | A02650 |
|---|---|---|---|---|---|---|---|
| MN | 55101 | Saint Paul | 44.95 | -93.09 | MN | 55101 | 13627.0 |
| MN | 55101 | Saint Paul | 44.95 | -93.09 | MN | 55101 | 40567.0 |
| MN | 55101 | Saint Paul | 44.95 | -93.09 | MN | 55101 | 55233.0 |
| MN | 55101 | Saint Paul | 44.95 | -93.09 | MN | 55101 | 42789.0 |
| MN | 55101 | Saint Paul | 44.95 | -93.09 | MN | 55101 | 72039.0 |

### Next, we needed to calculate median income.

Once we have a city attached to our income information, we must split-apply-combine our data to get a median income measurement for each city, state, and coordinate combination.

Before we do this, we must create two new columns: adj_lat and adj_lng. **The purpose of these is to allow for more specific income calculations, as there are usually multiple zip codes per city. Additionally, including these columns when merging will also account for some states having multiple cities with the same name.**

The Income and ZIP Code databases each have precise (but slightly different) latitude and longitude coordinates. Therefore, in order to create these adjusted lat/long columns, we truncated the original lat/long coordinates in each database.

**Example: Calculating Median Income For A City**

| primary_city | state | adj_lat | adj_lng | A02650 |
|---|---|---|---|---|
| Saint Paul | MN | 44.0 | -93.0 | 91710.5 |
| Saint Paul | MN | 44.0 | -92.0 | 155377.5 |
| Saint Paul | MN | 45.0 | -93.0 | 192374.0 |
| Saint Paul | MN | 45.0 | -92.0 | 34142.5 |

At this stage, we can finally merge our income information with our recruiting information as we have a **single record for each city, state and coordinate combination.**

# Is The Playing Field Really Level?

## Key Data Manipulation Steps

### Time to merge our income data with our recruiting data

**Recruiting Dataset**

| | year | ranking | committedTo | position | height | weight | city | hometownInfo.latitude | hometownInfo.longitude | lat_adj | lng_adj | stateProvince |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 2011 | 404.0 | Minnesota | offensive line | 76.0 | 280.0 | Saint Paul | 44.9504037 | -93.1015026 | 44.0 | -93.0 | MN |

**Income Dataset**

| primary_city | state | adj_lat | adj_lng | A02650 |
|---|---|---|---|---|
| Saint Paul | MN | 44.0 | -93.0 | 91710.5 |
| Saint Paul | MN | 44.0 | -92.0 | 155377.5 |
| Saint Paul | MN | 45.0 | -93.0 | 192374.0 |
| Saint Paul | MN | 45.0 | -92.0 | 34142.5 |

**Final Combined Dataset**

| | year | ranking | committedTo | position | height | weight | city | lat_adj | lng_adj | stateProvince | A02650 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 2011 | 404.0 | Minnesota | offensive line | 76.0 | 280.0 | Saint Paul | 44.0 | -93.0 | MN | 91710.5 |

### Additional Manipulation Steps

- Group recruits into more consistent positions bringing the count of unique positions down from **28** to **9**
- Drop international players from our dataset
- Drop universities that **are not in the top 100** in terms of number of recruits from 2011 - 2023
- Drop any duplicate recruits using names and unadjusted coordinates
- We also rename "**A02650**" to "**average_income**"

As you can see from the merged results, using the adjusted longitude and latitude coordinates allows for **more specific merging.** We cannot use unadjusted coordinates because the two datasets have **different initial coordinates.**

# Is The Playing Field Really Level?

## Analyzing The Origin Of Top Recruits

### Which States Have the Most Recruits?

The largest numbers of recruits tend to originate from states with either large populations or warmer weather (meaning longer football seasons). Texas, Florida and California have provided the most total prospects since 2011. Regionally, the Southeastern U.S. features 5 of the top 10 sources of recruits (Florida and Georgia, plus Louisiana (#6), Alabama (#7), and Tennessee (#9)). These 5 Southeastern states in the Top 10 have produced ~31% of all recruits since 2011.

We also show the percentage of a state's recruits that receive 5-star ratings and the number of top recruits per million people. These are useful proxies for the perceived quality of a state's players. Recall that 5-star prospects only represent the top ~1% of all recruits. Florida has the highest proportion of 5-star recruits while Georgia has the most 5-star recruits per million.

### Regional Considerations

As expected, the map on the right shows higher concentrations of prospects from regions with a higher population density (the East Coast, Ohio, California) or better weather (Texas, the Southeast).

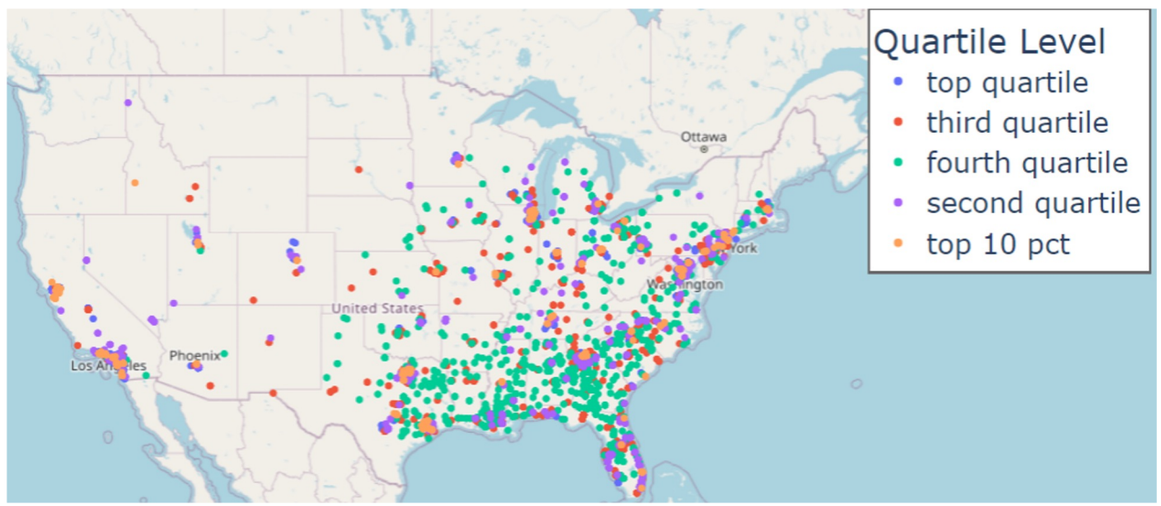
But importantly, college recruiting tends to exhibit a local bias. Many high school players choose to play closer to home if possible. This suggests that colleges located in wealthier areas (typically large coastal cities) would be expected to recruit players from higher income quartiles on average, while colleges in less affluent areas will tend to recruit from poorer income quartiles.

### Largest Number Of Recruits By State: 2011-23

| | Stars | | | Total | % of Total | 5-Star % Of | Population | Recruits Per MM | |
|---|---|---|---|---|---|---|---|---|---|
| State | 3 | 4 | 5 | Recruits | Recruits | State Total | (MM) | 3-5 Star | 4-5 Star |
| Texas | 3,235 | 600 | 63 | 5,782 | 13.4% | 1.09% | 29.5 | 132 | 22 |
| Florida | 3,198 | 597 | 81 | 5,635 | 13.1% | 1.44% | 21.8 | 178 | 31 |
| California | 2,272 | 448 | 44 | 4,291 | 10.0% | 1.03% | 39.2 | 70 | 13 |
| Georgia | 1,987 | 384 | 47 | 3,402 | 7.9% | 1.38% | 10.8 | 224 | 40 |
| Ohio | 1,026 | 173 | 12 | 1,880 | 4.4% | 0.64% | 11.8 | 103 | 16 |

Note: Total Recruits column also include 1 and 2-star recruits.

### Hometowns and Income Quartiles Of 5 and 4-Star Recruits (2011-23)




Note: Top Quartile excludes recruits in the Top 10 Pct.

# Is The Playing Field Really Level?

## Analyzing Individual Colleges: Case Study – Georgia vs. UCLA

**The tendency to recruit close to home significantly influences the overall income distribution of a school's prospects. The tables below show how each school's prospects are distributed by quartile. If the distribution was completely balanced, we would expect ~25% of the prospects to be in each quartile and ~10% to be in the top decile. In order to analyze numerous schools and identify patterns, we developed a function that allowed us to make quick comparisons between any given school(s).**

### Georgia (5 and 4-star prospects: 2011-23)

- Significant component of recruiting classes from Georgia or nearby
- Heavily weighted towards bottom half of income distribution – 32% from Fourth Quartile and another 21% from Third Quartile
- Top Quartile (18%) and Top Decile (5%) both underrepresented
- Despite this apparent imbalance, roughly in-line with SEC averages
- Local recruiting base bolstered with elite prospects from across the country

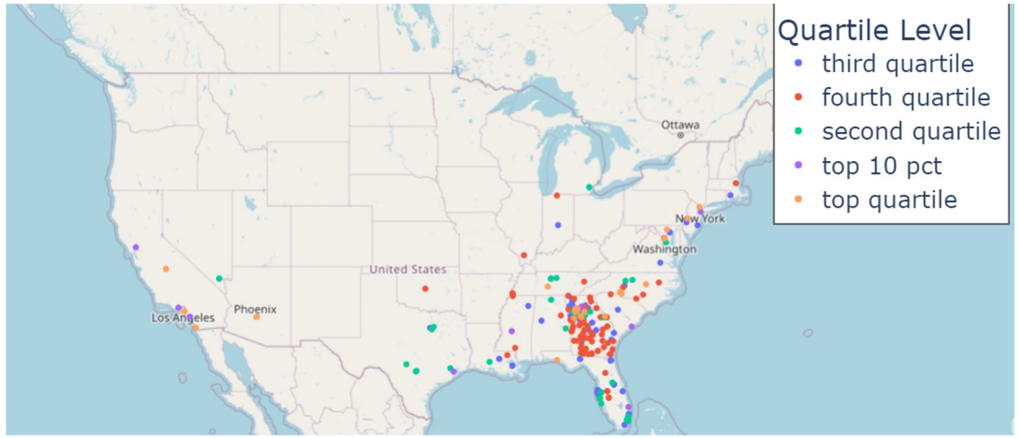
| | Q4 | Q3 | Q2 | Q1 | Top 10% | Total |
|---|---|---|---|---|---|---|
| Recruits | 67 | 44 | 62 | 37 | 10 | 210 |
| % of Total | 32% | 21% | 30% | 18% | 5% | 100% |
| SEC Avg. | 33% | 27% | 21% | 19% | 8% | 100% |

### UCLA (5 and 4-star prospects: 2011-23)

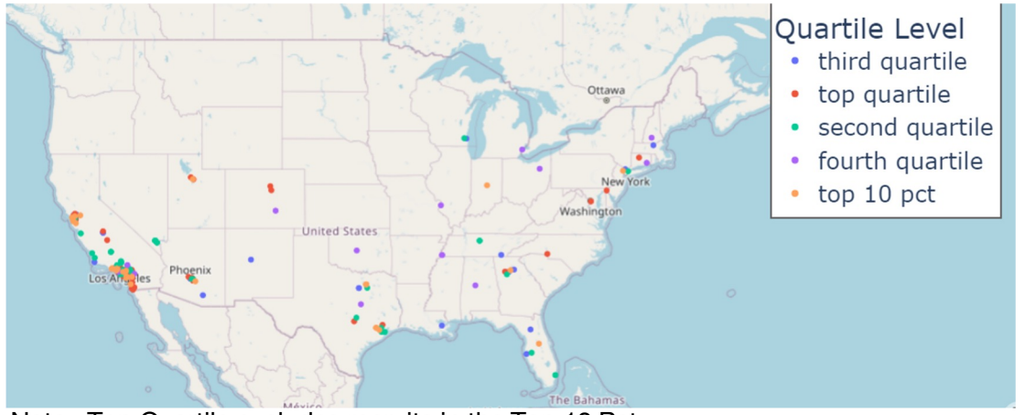
- Recruiting weighted toward Southern California, where income levels are higher than national averages. In-line with Pac-12 averages.
- 51% of total prospects from Top Quartile of income distribution
- Fourth (8%) and Third (13%) Quartiles underrepresented

| | Q4 | Q3 | Q2 | Q1 | Top 10% | Total |
|---|---|---|---|---|---|---|
| Recruits | 7 | 12 | 25 | 45 | 17 | 89 |
| % of Total | 8% | 13% | 28% | 51% | 19% | 100% |
| Pac-12 Avg. | 7% | 20% | 27% | 46% | 18% | 100% |

### Hometowns of Georgia Recruits: 2011-23




### Hometowns of UCLA Recruits : 2011-23




Note: Top Quartile excludes recruits in the Top 10 Pct.

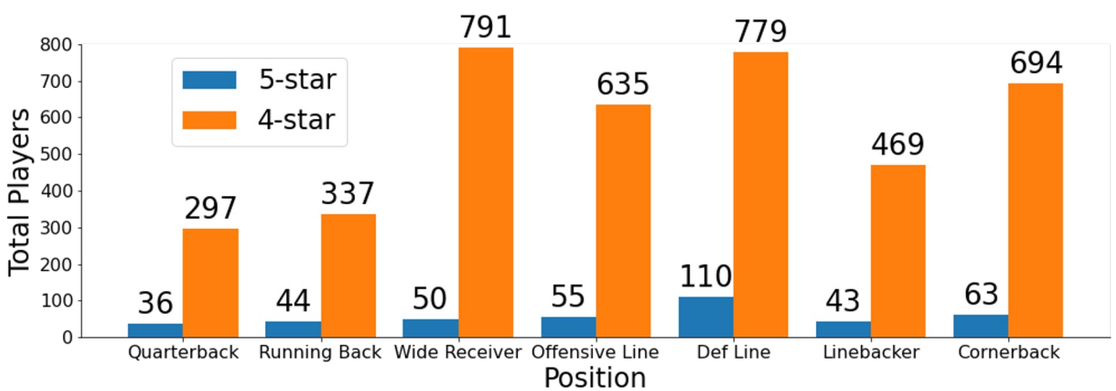
# Is The Playing Field Really Level?

## Analyzing Aggregate Prospect Distribution

### Summary Of Star Ratings By Position

Tabulating the number of prospects by position and star rating and showing the data in a grouped bar chart provides a helpful framework for understanding the numbers of elite prospects. For instance, there were only 36 total 5-star quarterbacks and 296 4-star quarterbacks from 2011-23. Given the relatively small number of 5-star players (average of 32 per year), we have generally treated both 5 and 4-star players as 'premium' recruits. It is also worth noting that there are varying numbers of players for each position group on the field at any given time – typically 5 Offensive Linemen, 2-3 Wide Receivers, 3-4 Defensive Linemen, 3 Linebackers and 4 Cornerbacks. This explains the wide differences in the total counts by position.

### Total 5 & 4-Star Recruits By Position: 2011-23




### Distribution Of Top Prospects By Income Quartile

In order to better understand the relationship between star rankings and the U.S. income distribution, we have narrowed our focus to 5 and 4-star rated players (roughly the top 11% of all prospects).

We then calculated the percentages of players from 2011-2023 whose hometowns are in each income quartile (plus the top 10%). Much like our Georgia vs. UCLA analysis, we would expect ~25% of all prospects to originate from each quartile and ~10% to originate from the Top Decile if the distribution was completely balanced.

| | Income Quartile | | | | |
|---|---|---|---|---|---|
| Stars | Q4 | Q3 | Q2 | Q1 | Top 10% |
| 5 | 24% | 25% | 27% | 24% | 10% |
| 4 | 23% | 25% | 25% | 27% | 11% |

Note: Includes recruits from 2011-23.

A few notable observations:

- At the **aggregate level** (across all position groups), the distribution is reasonably balanced across the quartiles
- For **5-star prospects**, there are slightly (~2%) more prospects in the Second Quartile than expected
- For **4-star prospects**, the Top Quartile is overrepresented by ~2%

# Is The Playing Field Really Level?

## Analyzing Position-Level Distribution Of Prospects By Income Quartile

**Breaking down the aggregate data by position using a groupby function shows much wider variations in the percentage of players from each income quartile – most notably among Quarterbacks and Offensive Linemen.**

### Offense

| | | Income Quartile | | | | |
|---|---|---|---|---|---|---|
| **Position** | **Stars** | **Q4** | **Q3** | **Q2** | **Q1** | **Top 10%** |
| Quarterback | 5 | 17% | 19% | 25% | 39% | 14% |
| | 4 | 18% | 24% | 21% | 37% | 18% |
| Running Back | 5 | 27% | 25% | 25% | 23% | 9% |
| | 4 | 28% | 25% | 24% | 23% | 8% |
| Wide Receiver | 5 | 22% | 28% | 26% | 24% | 8% |
| | 4 | 21% | 22% | 29% | 28% | 11% |
| Offensive Line | 5 | 15% | 24% | 25% | 36% | 16% |
| | 4 | 22% | 24% | 23% | 31% | 13% |

### Defense

| | | Income Quartile | | | | |
|---|---|---|---|---|---|---|
| **Position** | **Stars** | **Q4** | **Q3** | **Q2** | **Q1** | **Top 10%** |
| Defensive Line | 5 | 30% | 26% | 26% | 18% | 9% |
| | 4 | 26% | 27% | 24% | 23% | 10% |
| Linebacker | 5 | 28% | 28% | 28% | 16% | 12% |
| | 4 | 23% | 27% | 26% | 24% | 11% |
| Cornerback | 5 | 17% | 26% | 33% | 24% | 10% |
| | 4 | 22% | 25% | 27% | 26% | 11% |

### Key Observations:

- **Quarterbacks:** Largest imbalance toward the Top Quartile of any position. For example, 39% of 5-star QBs are from locations in the Top Quartile, while 14% are from hometowns in the Top Decile.
- **Running Backs:** Modest underrepresentation in the Top Quartile (23%) and overrepresentation in the Bottom Quartile
- **Wide Receiver:** Modest bias toward Top and Second Quartiles among 4-star receivers
- **Offensive Line:** Heaviest skew towards the Top Quartile of any position other than Quarterback
- **Defensive Line:** Distribution skews away from Top Quartile and toward Bottom Quartile.
- **Linebackers:** 5-star recruits biased toward the bottom 3 Quartiles. 4-star recruits more balanced
- **Cornerbacks:** Biased toward Second Quartile

### Statistical Significance:

We also tested whether the results for each position were statistically different from the overall distribution of recruits. Using a Mann-Whitney U test, we found that at an alpha of 0.05 the average income of prospects is significantly different from the rest of the recruit population for all position groups except Offensive Linemen and Linebackers (and for all groups at an alpha of 0.10).

# Is The Playing Field Really Level?

## Analysis – Performance Impact & Analytical Diversions

**A natural question that arises from our analysis is whether higher star rankings ultimately impact performance. If wealthier prospects do get disproportionate numbers of high star rankings, does that translate to on-field individual (and team) performance? If, for example, a large share of 5-star QBs are from wealthier zip codes but underperform, it might suggest that the 5-star rankings were misallocated.**

### Correlation Analysis

There is a significant positive statistical relationship (0.586) between the **average star ranking** of a college's recruiting class and the **team's winning percentage**. Results are very similar for "rating", a proprietary metric for each player in the college football database.

At the **position level**, there are **weaker but still positive statistical relationships** (~0.18 - 0.28) between average star ranking and team winning percentage, but these are still directionally correct for each position.

**Star Ranking vs. Winning %**

| | win% | rating | stars |
|---|---|---|---|
| **win%** | 1.000000 | 0.586174 | 0.586701 |
| **rating** | 0.586174 | 1.000000 | 0.996709 |
| **stars** | 0.586701 | 0.996709 | 1.000000 |

**Position vs. Winning %**

| | position | variable | win% |
|---|---|---|---|
| **35** | special teams | stars | 0.178222 |
| **23** | quarterback | stars | 0.217516 |
| **19** | other | stars | 0.223605 |
| **27** | receiver | stars | 0.250262 |
| **31** | secondary | stars | 0.260744 |
| **11** | linebackers | stars | 0.267615 |
| **15** | offensive line | stars | 0.275081 |
| **3** | backs | stars | 0.277587 |
| **7** | defensive line | stars | 0.284801 |

### Correlation Analysis (continued)

We can also look at the relationship between **star ranking** and **on-field productivity**. Metrics such as Predicted Points Added (PPA) show how much better a player performs over expectations for a given situation. In this example, **QBs with higher star rankings tend to be modestly more productive (coefficients of ~0.18 – 0.20).**

**Star Ranking vs. QB Productivity**

| | variable | stars |
|---|---|---|
| **3** | averagePPA.all | 0.170708 |
| **11** | totalPPA.all | 0.185317 |
| **12** | totalPPA.pass | 0.183665 |
| **17** | totalPPA.standardDowns | 0.197421 |

### Analytical Wild Goose Chases

Our analysis of the combined data sources took us in many directions. All of these did not end up being fruitful. Some examples:

- **Average Income Per School Over Time:** We had hoped to identify trends in specific schools (or between groups of schools) and created line charts to do so. Unfortunately, we did not find trends that were meaningful enough to warrant inclusion.
- **Quartile Outliers**: We extensively debated how to measure the largest outlier favoring any given quartile. The goal was to identify schools which strongly favored each of the four quartiles, but once we realized the importance of regional factors and the relatively balanced nature of the distribution at the aggregate level, we abandoned this idea.

# Is The Playing Field Really Level?

## Conclusions & Opportunities

### Summary of Key Findings:

Based on our research we came to the following conclusions:

- At the **aggregate level**, the distribution of 5 and 4-star prospects is **reasonably balanced** across the income distribution.
- Recruiting is greatly influenced by **regional factors.** Many players prefer to play closer to their hometowns. Recruiting classes at individual schools tend to reflect the income profile of their locations.
- A more granular analysis at the **position level** reveals **much wider discrepancies** in the distribution of players across income quartiles. These differences are statistically significant for most positions.
    - **Quarterbacks** and **Offensive Lineman** skew most heavily toward the Top Quartile. The perceived importance of the QB position to team success has led to the emergence of private training sessions and specialized training camps at the youth level. These can be expensive and are not available to players from all income quartiles. This suggests that **the playing field is not entirely level across all positions**.
    - **Running Backs** and **Defensive Lineman** skew most heavily toward the Fourth Quartile.
- Correlation analysis suggests that 5 and 4–star prospects have a positive – albeit mild -- impact on a team's winning percentage relative to 3, 2 and 1-star prospects. These prospects are also more productive than lower ranked players based on several on-field metrics.

### Opportunities For Further Research:

There are other areas of research that are beyond the scope of this project that would warrant further consideration:

- **NFL Success:** Our analysis focused primarily on the link between high school star ratings and income distribution. We did not examine how the rankings of high school players from various income brackets correlates to success at the NFL level. Is a 5-star rating a good predictor of NFL success? Does it vary by position?
- **Regional Biases:** Is there a significant difference in on-field performance between 5-star recruits from different regions? Do recruits from perceived football "hot beds" such as the Southeast tend to perform well at the college level? Or is it a case of "regional reputation" influencing the star rankings assigned by high school scouts?
- **Physical Measurables:** How much do physical traits (height, weight, BMI, hand size for Quarterbacks, etc.) influence star rankings? Most football positions tend to have accepted ranges for these metrics, but does the old adage of "He looks like a football player" still apply? Are there also potential biases (based on race, region or income level) within the evaluations of scouts who create the star ratings?
- **Relative Valuation Of Positions:** Are position groups appropriately valued based on each's group's individual contribution to winning percentage? Our preliminary analysis suggests that Offensive and Defensive Lineman make the greatest contribution to team winning percentage. Are high school scouts and colleges both mistakenly focusing on higher-profile positions such as Quarterback rather than more impactful ones?

# Is The Playing Field Really Level?

## Statement of Work & References

### Process Overview:

Our general approach for this project was to maintain an active, collaborative dialogue as much as possible. Once we initially combined our data sources, we each explored different ways to analyze the dataset in search of evidence supporting our working thesis. We then shared findings via Slack or on our periodic check-in calls and decided as a group how to further refine the analysis. In the future, one area for improvement might have been to have an initial group brainstorming session where we more formally defined research subtopics for each of us. This might have increased the efficiency of the process – at least in the initial phase.

### Statement of Work:

### Greg Holden:

Background in corporate finance and capital markets. Extensive experience analyzing data and crafting narratives around potential financing or strategic transactions. Focused on the generation of analytical ideas, searching our combined data sources for trends, creating visualizations, and the presentation of our collective findings.

### Jacob Schmidt:

Background in Econometrics and Data Science. The majority of my professional experience has been in predictive model building, with a focus on natural language tasks. Based on my experience I focused on data manipulation, feature engineering, correlation analysis and any ad hoc tasks necessary.

### Tawfiq Zureiq:

Background in Software Engineering and Data Science. Most of my experience has been in Development and Data Analytics. I focused on data manipulation and visualizations.

### References:


Bergman and Logan (2020). "Revenue Per Quality of College Football Recruit". *Journal of Sports Economics*, 21(6):152700252092122. Retrieved from https://www.researchgate.net/publication/341411022_Revenue_per_Quality_of_College_Football_Recruit

Mankin, Rivas & Jewell. (n.d). "The effectiveness of college football recruiting ratings in predicting team success: a longitudinal study". *Research in Business and Economics Journal,* Vol 14, 1-19. Retrieved from http://www.aabri.com/manuscripts/182841.pdf

McLaughlin, Bryce, "College Football Recruiting and the Correlation to Success" (2021). Honors Theses. 413. https://digitalcommons.coastal.edu/honors-theses/413

Wheeler, Nicholas, "Do High School Football Recruit Ratings Accurately Predict NFL Success?" (2018). *CMC Senior Theses*. 1846. https://scholarship.claremont.edu/cmc_theses/1846

Wittry, A. (2019, July 2) "Analyzing College Football's Relationship Between Recruiting Class Rankings and Wins". Watchstadium.com. https://watchstadium.com/analyzing-college-footballs-relationship-between-recruiting-class-rankings-and-wins-07-01-2019/