Week 3 | Team Project Milestone Report

Bradley Jones, Lucas Sofolo, Mario Lloyd Galvao-Wilson,

Our team would like to aid collegiate football recruiters in their search for talent, by providing recommendations for recruiting areas and specific builds based on current and former player statics and attributes. We plan to identify the hotbed areas in the United States for NFL talent to suggest as youth training programs and possible outreach programs. This information is not limited to recruiting, but can also aid in team expansion to increase fans in specific areas.

Football scouts of all levels can benefit from this information as they can get a head start on the development and recruitment. Leagues (collegiate and professional) can also benefit as they can invest more into areas with growing interest and also introduce the game to those unfamiliar. We hope to discover in our analysis what levels benefit most from this data.

Data inspection activities should give one a general overview of the dataset and its dependability by checking for issues like missing data or unclear output. Data visualization allows one to notice more trends, better understand, and communicate findings about the data.

The data is thoroughly inspected, imported, and processed in the right manner. It was imported without errors or messy output and processed following all proper exploratory steps. The exploratory stage helped us determine if missing values exist, data is readable, and columns represent what they should be through functions head() and info(). Every step is completed to fix data, including correcting the height column and increasing the dependability of the height data with a definition. Data visualization steps are also started with a scatterplot showing player heights for each position and how they compare.

Thus far, our team has used Pandas multiple times because of its extensive number of helpful tools for cleaning and analyzing datasets. Some we have used include read_csv(), head(), info(), unique(), map(), isna(), and apply(). Read_csv() is present in the first true line of code to retrieve data from the NFL file and output it as a data frame. Next, we used head()

to return the top 5 values and info() to get a summary. The functions were a quick way for our team to gain an overall understanding of the dataset.

Another Pandas function we used was unique() to return each specific height represented in the column. The height column had incorrect date data, so we applied map() to substitute the wrong values of the column with the corrected version called date_to_height. Isna() is incorporated when converting height to inches to detect missing height values and apply() to convert all heights to inches. We also used NumPy for mathematical equations such as sum() to count missing values in each category and provide insight into which missing data is most prevalent.

Some key insights we have discovered from our first scatter plot involving height is that offensive tackles (OT) in the NFL have the tallest low height of any position at 73 inches and highest at just under 82.5. The statistic makes sense as long arms are one of the most sought-after physical attributes in tackles because it allows them to reach out and grab the chest of a defensive end before they can get around them. While our insight isn't new, it could be useful in verifying that short-armed tackles don't make it in the NFL.

A college football recruiter could learn that tackles shorter than 73 inches usually have shorter arms and are less likely to be successful at that position. As a result, when recruiting for tackles, they could focus only on players taller than that, ultimately increasing their chance of signing a successful tackle.