NFL Exploratory Data Analysis

Ismael Saleem

Big Data Project Report

Abstract

For my project, I participated in a Kaggle competition mainly focusing on Exploratory Data analysis. The competition was titled “NFL Big Data Bowl”. The data I dealt with consisted of 7 CSV files. These files charted data regarding NFL players, different plays, games, tracking the players, and scouting. The competition was very unrestricted in nature and allowed for creativity in what insights should be gained and how to explore the data.

Although the vastness of the data may have been overwhelming, by using a combination of Pyspark and Jupyter Notebook, I was able to gain several insights from data provided. I was able to successfully recreate plays, find relations between different player stats and their positions, and gain insight into the plays and possession throughout the years 2018 and 2020.

Keywords

Big Data, Kaggle, Data Analytics, NFL

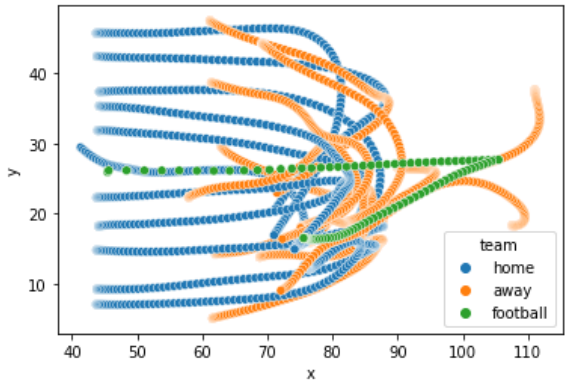
Problem Statement

The main questions I sought out to answer in this project were:

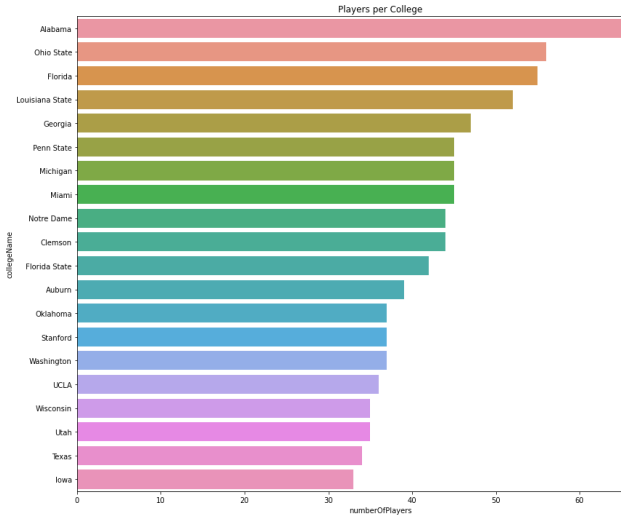
- Which software/tools would be best for analysis
- Is it possible to visualize the plays
- Is there any variation or relation between weight/height and the position a player plays
- Which colleges produced the most NFL athletes in that time period
- What was the most common type of play in this time period

Approach

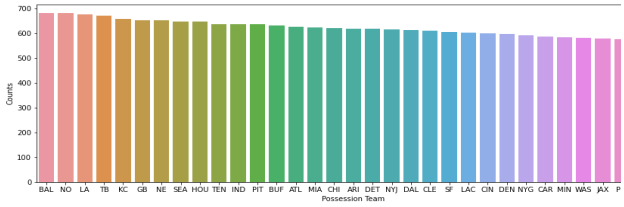
As stated previously, the dataset provided in this kaggle competition consisted of 7 CSV files containing 113 columns overall and taking up around 5 GB of space. The “players” file contained stats for the different NFL players that year (including their height, weight, date of birth, etc). The “plays” and “games” file contained data dealing with the type of plays done in different games and quarters. The tracking files for years 2018 to 2020 actually tracked where specific players were in relation to the field during a particular play.This was done by giving each player an X and Y axis representing their location. When considering all of the data and what insights could be derived from it, it was obvious that the best way of analyzing it would be through exploratory data analysis. To do this, I knew I wanted to use a python based tool as opposed to SQL, as python is very effective for data visualization. At first I tried to use Google Colab, however the datasets (specifically the tracking files) were too large for the system to handle. So I ended up using a bit of Pyspark as well as Jupter notebook for visualization as these two tools were able to process the larger datasets with no issue. One of the main goals I had was to successfully be able to use the tracking data to simulate specific plays. After some consideration, it was clear that the best way to do this would be by using a scatter plot, each dot representing a player’s movement through the field or the ball’s trajectory after it’s been kicked.



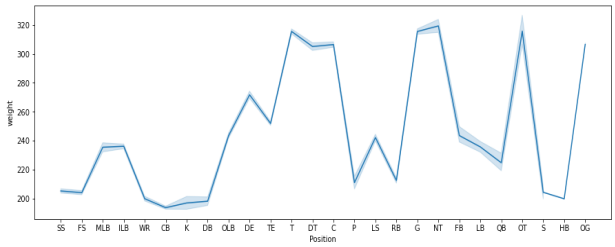
In terms of some of the other analysis goals, I was able to draw all of the insights that I set out to look into. By gathering the value count for each mention of the different colleges in the “players” dataset, I was able to find that Alabama had the most amount of graduates that ended up in the NFL between 2018 and 2020

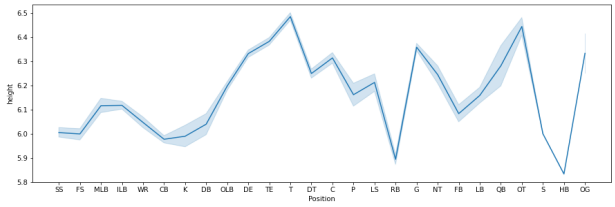


I was also able to determine that the Baltimore Raven (named below as BAL) had the most amount of possessions out of any team in that time period

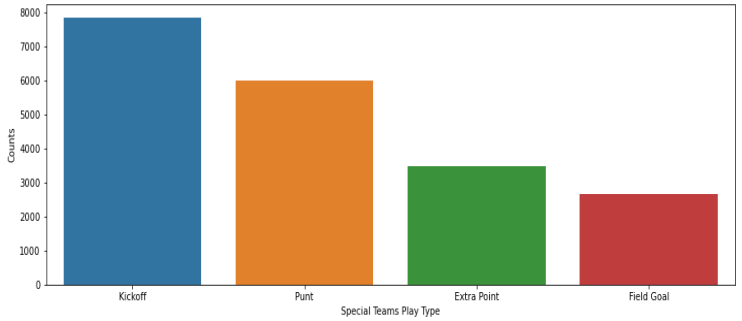


In terms of data with regards to weight and height, the average weight of an NFL player between 2018 and 2020 was 244 pounds and the average height was about 6 foot 2 inches. With that being said, both weight and height varied quite a bit between different positions.





Finally, one of the main insights I wanted to gain was what was the most common type of play throughout the year 2018 to 2020. By gathering the counts of different plays throughout those years, it was clear a kickoff was the most common type.



Conclusion

Using both Pyspark and Jupyter Notebook, I was successful in gaining the insights I set out to gain using the datasets provided. As the competition was very much open ended, there was much to explore. The main challenge was choosing the appropriate tools that can handle such a large data set, as well as choosing queries

that can clean and visualize the data. At the end, with the right tools, I was able explore almost every aspect that could prove useful.

References

https://www.kaggle.com/competitions/nfl-big-data-bowl-2022