CSE 314A Team 1 Project Proposal

Group Infrastructure:
Our group wants to find insights in sports data, namely football data. Gus is interested in working with a football dataset, specifically the NFL Big Data Bowl's defensive dataset which rolled out in 2021. Similar to Gus, Matt is going to explore football data sets, mainly the NFL Big Data Bowl dataset, to try to answer an exploratory question. While not set on a question, an example could be what down is the least successful for blitzing. Ram is interested in looking at the data in a economic standpoint (what big contracts were busts and what weren't) and if there is a possible model that returns a reasonable contract size given statistics. Hannah is interested in the effect of college play on pro athletes, maybe looking at whether a certain school's athletes adjust faster/better to pro level or whether a certain school is especially strong in a certain position.

Background:
Right now, a lot of existing data sets are pure stats or advanced stats about players and do not offer us a way to analyze or get insights from the data or lack situation specific measures. As we watch sports, we have different questions, such as which university produces the best athletes, that is not easily answered by looking at raw statistics like those on Sports Reference.

Scope:

Hannah will focus on one specific university or a small group of universities and the performance of their former players who have turned pro. Ram will focus on NFL players and will possibly feed a dataset into a learning model. Matt will focus on using the NFL Big Data Bowl's defensive dataset and possibly the sports reference football datasets to evaluate a specific research question that will be decided. Gus will focus on defensive statistics, calculating the performance of the defensive players, whether that statistic already exists or I create a new one to evaluate performance.

Outline:
We will start with identifying all of the data sources that we might need. For example, the NFL big data bowl data and Sports Reference will provide us with basic statistics and advanced statistics. Websites like Spotrac will provide us with contract information. We will figure out how to web scrape some of these websites for data. We will then individually identify what data we need to answer our own personal questions and do some analysis. We will then combine the different data sources and display the data on a dashboard. We will split up the tasks (data aggregation, coding the dashboard, creating graphs and visuals).

There are all kinds of sports data available for us and different operations that we could do to the data. The data is also found through different websites and APIs, so we would need to work on merging all the data sources and classifying the data properly.

Relevance:

As for the relevance to the course, the main group work we will be doing is wrangling and visualizing a lot of the data we are looking at for our respective questions. We are thinking about creating a dashboard similar to the type of dashboard we created in homework 2 where we can visualize data based on fields we input (like radio buttons or checklists) so we can all use the same dashboard to explore the data we're looking at. Aggregating, wrangling, and visualizing the data as a group is what this class is all about and we believe we will be doing a lot of each.

Potential Datasets:
https://www.kaggle.com/c/nfl-big-data-bowl-2021/data

https://www.kaggle.com/henggaocai/blitz-project-nfl-big-data-bowl-2021