To extract the NFL passing, rushing, and receiving stats, https://www.pro-football-reference.com was utilized. The “pd.read\_html” function in pandas was used to pull the applicable tables down from the site for every season spanning from 2009-2018.

Some cleaning measures that needed to be taken were the following:

1. Within the tables, players that had an asterisk next to their name in the player column had been selected to Pro Bowl, and players that had plus sign next to their name had been selected First Team All-Pro. To avoid future querying issues, the “str.replace” function was used within pandas to change those two characters to blanks.
2. Another issue with the original scraped tables is that they included additional column headers within the table at 30 player intervals to make is easier on a user that would be scrolling through the web page. These were removed be setting the dataframe to a new variable that dropped all of the rows that were simply a duplicate of the column names.
3. To allow querying to be performed on the specific year that the stats related to, a “Year” column was added in pandas.

Once the tables for the 10 years examined were scraped and cleaned, “pd.concat” was utilized to combine all of the dataframes into a consolidated dataframe. This was done 3 times to create 3 final tables (passing, rushing, and receiving). Columns that were not of interest for our querying in SQL were dropped from the final dataframes. Additionally, column headers that were capitalized needed to be cleaned and changed to lowercase. In SQL, when column headers are capitalized, querying becomes case sensitive by default and requires double quotes to be put around the column header when calling specific column queries. The change to lowercase was done to avoid this.

The second round of dataframes came from downloaded csv files. These data sets included personal information on all active players (such as height, weight, age, etc) along with salary information for all active players. The cleaning measures taken on the former dataframe included: 1) dropping columns that would be available through other tables in the database like “team” and “player id”. 2) Setting the “player” column to the index as it was going to be used as our key to connect the tables. 3) Setting the headings to lowercase as to make them more compatible with SQL.

The latter data set took more cleaning to be prepared for integration into the database. These steps included 1) changing the formatting of the player’s name so it would match the other data sets. Initially, the csv file had the names listed LastName, FirstName in one column. So, we used the string split function to parse the name and remove the comma. Once that was done, we put the players names back into the “player” column with the correct format. 2) We dropped retired players out of the data frame because this specific one had players from as far back as 1953. They would not have matched up with any of the information we had, so were dropped. 3) Dropped columns that were not useful to the dataset or had many missing values, as many of the rows, like “High School”, were mostly incomplete.

Once all 5 tables were pulled into pandas and cleaned, a connection to SQL was established using SQLalchemy and all tables were populated in SQL. This connection allowed us to use a relational database to perform joins on specific players to obtain their college, various physical attributes, and salary information for specific years. Since many people are interested in certain players, we decided the best way to go about joining most of the data was on player names.

There was a thought of using some sort of unique id, however the data sets we pulled from all had differing player id’s, and one was not readily available from the NFL or NFLPA to use for our database.

These datasets were chosen to create a readily available database of players that could be easily queried to find out information. For example, if one wanted to find out all of the current NFL quarterbacks that played at Clemson and how they have fared statistically in the league, this database would be a useful tool.