**ETL Project Final Report**

Brent Thomas

Emmanuel Olofinkua

Matt Houser

*February, 29 2020*

**Extract**

For this project we are asked to create and perform ETL successfully on multiple sources of data. We did this by extracted data from two separate websites, bleacherreport.com and USAtoday.com. These websites contained articles that discussed the all-time NFL draft picks, more specifically the best and worst picks for each team. The bleacher report article discussed first round picks only while the USA today article included all rounds.

Each team’s best and worst NFL draft picks (1st round only)

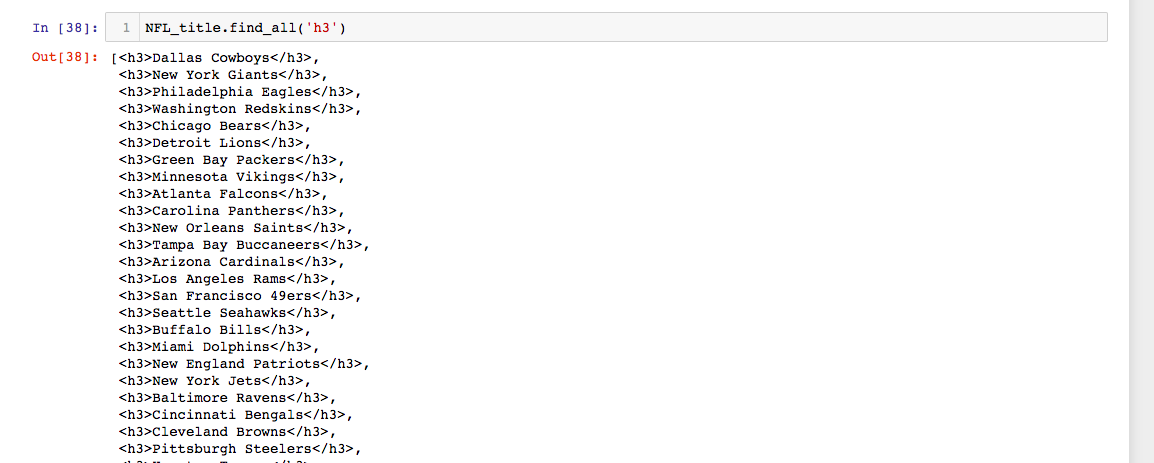
<https://bleacherreport.com/articles/2767040-every-nfl-teams-best-and-worst-1st-round-draft-pick-of-the-super-bowl-era#slide4>

Each team’s best and worst NFL draft picks (all rounds)

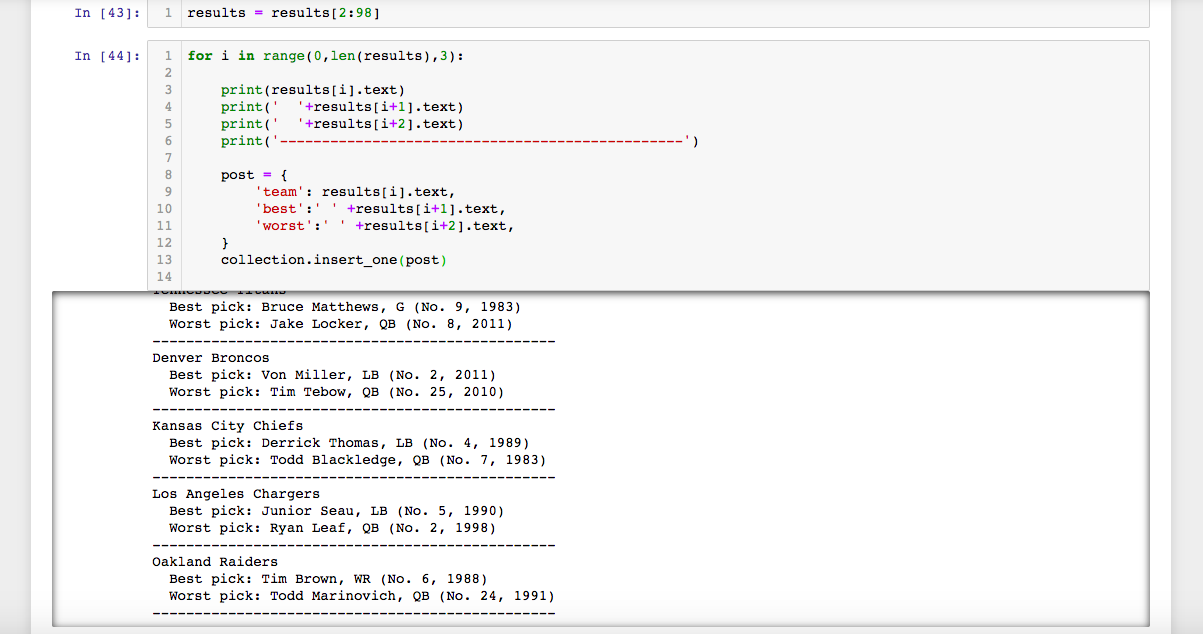
<https://ftw.usatoday.com/2018/04/nfl-best-worst-picks-team>

**Transform**

To clean and transform our datasets, within Jupyter Notebook we first created a path to open chrome driver and defined the URL. We then used find.all to display all the paragraphs, titles, and headers within each article for reference.



Using a for-loop we were able to print these references to narrow down results. We then used the import re (regular expression) function to include only the teams with their corresponding best and worst draft picks. After this, we ran an additional for-loop that would loop through all three pieces of data (team, best pick, worst pick) and separate them for loading purposes.



When scraping the Bleacher Report article, we came across an issue with a blank string interrupting the data, causing an extra space between players and throwing the data set off. We needed to filter the blank lines out that were occurring around when there was a break in the html.

**Load**

After we scraped our web pages and loaded them into the data frames, we created a local connection to mongoDB and loaded both collections to our NFL\_Best\_Worst\_db in Mongo.