Kyle Davis

5-18-2019

ETL Project

**Abstract**

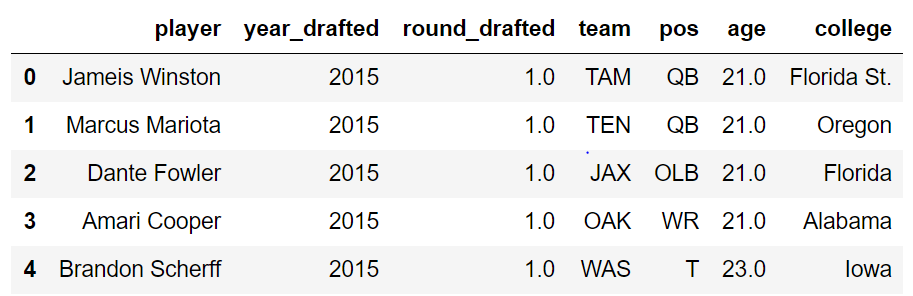
To compile a relational database of NFL player data, including draft position, college attended, team drafted, position played, arrest incidence, and suspension by the league. The goal of this database might be to inform a team’s yearly draft board. For example: if many players drafted in the later rounds (rounds 4-7) from a particular school are more prone to adverse behavior, you might want to consider a similar project player from another school less prone to arrests instead, all else being equal.

**Extract**

The process for our extraction involved downloading datasets from Kaggle in CSV form. These included NFL player draft data, player arrest data, and player suspension data. Original data was sourced from pro-football-reference.com and, fivethirtyeight.com, and the work of a USA Today reporter Brent Schrotenboer.

**Transform**

First, I created pandas dataframes of the datasets that I was working with, excluding irrelevant info (sacks, tackles, yards, etc.).

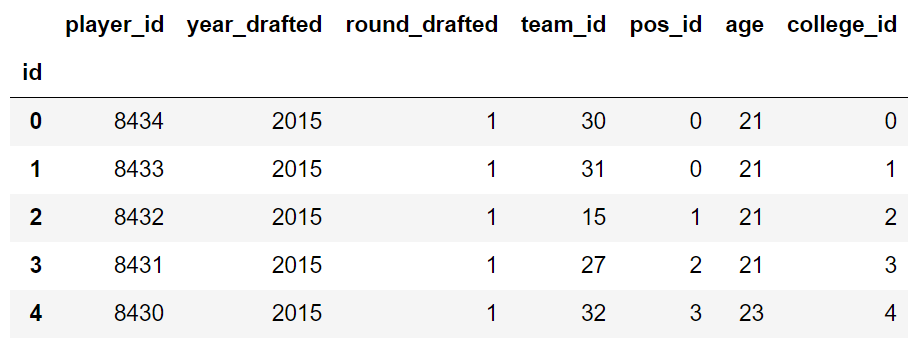


Then I standardized the team names using standard broadcast tags, and standardized position abbreviations. Examples include: JAX and JAC both being used to represent Jacksonville, and each dataset using different names for free agents (FREE vs. Free agent)

Next I created several indexed dataframes in python to be able to upload into a MySQL server. For example, here is the finished team abbreviation dataframe:



Using these dataframes, I replaced the strings in each of the original dataframes with the index of the item in each.

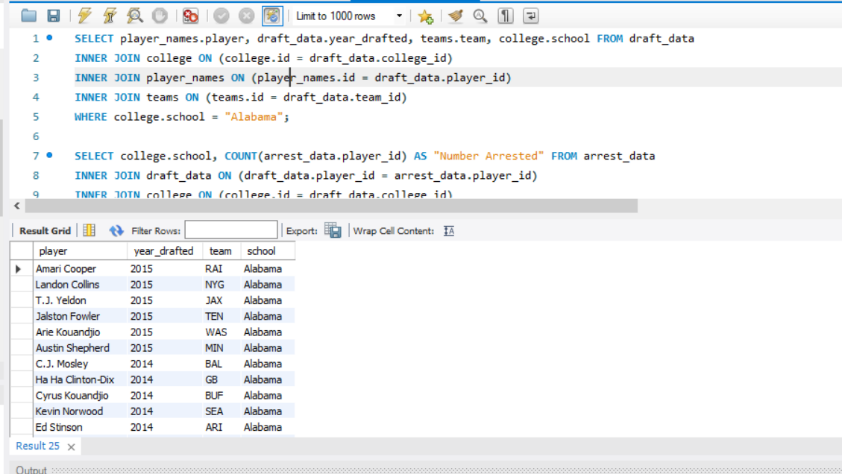


I attempted to do this transform on the player suspension data, but I came across a bit of trouble when doing so due to the player names being abbreviated. Had I had more time with this project, I may have been able to come up with a way to approximate the names, but the main issue was that there may be 15 players drafted with the last name “Jackson” and with a first initial of “B.” Because of this difficulty, due to time constraints, I decided to only work with the arrest and draft data.

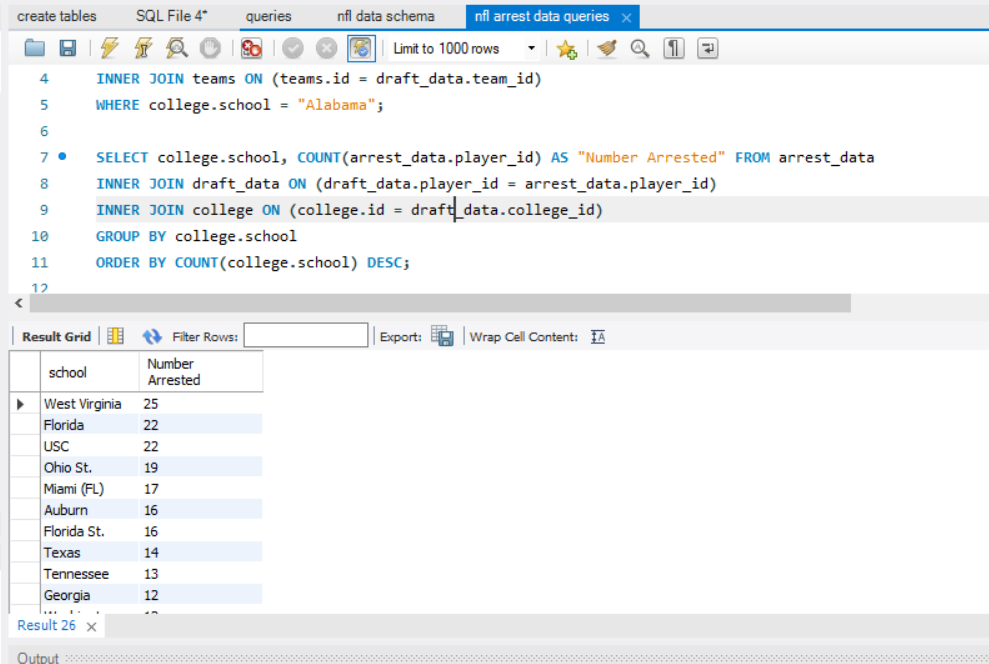
**Load**

The load process was quite simple. I created a connection to a MySQL server and created a schema there to apply the transformed data to. To confirm that everything was working properly, I performed a few queries:

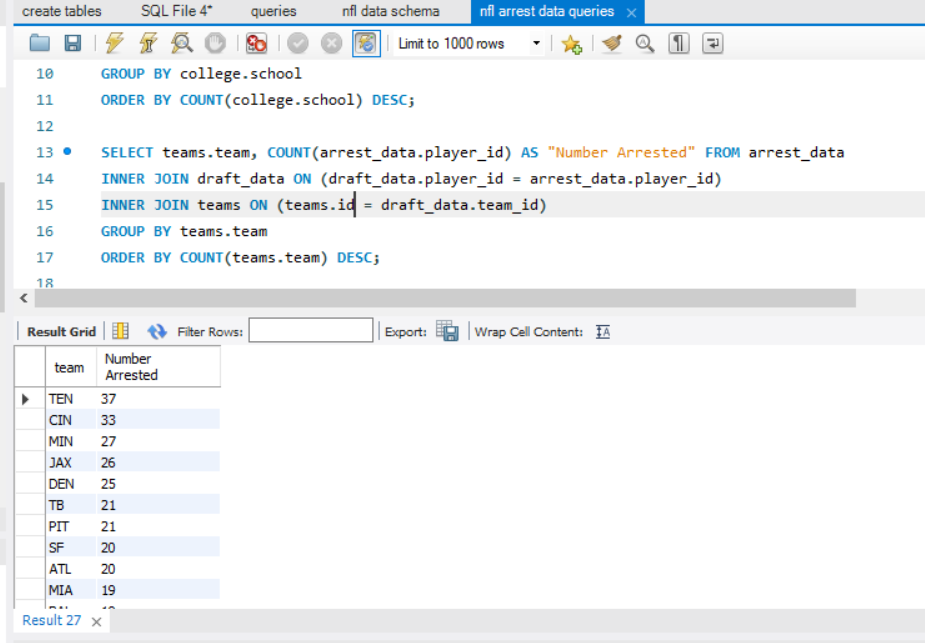
1. Players drafted by Alamaba



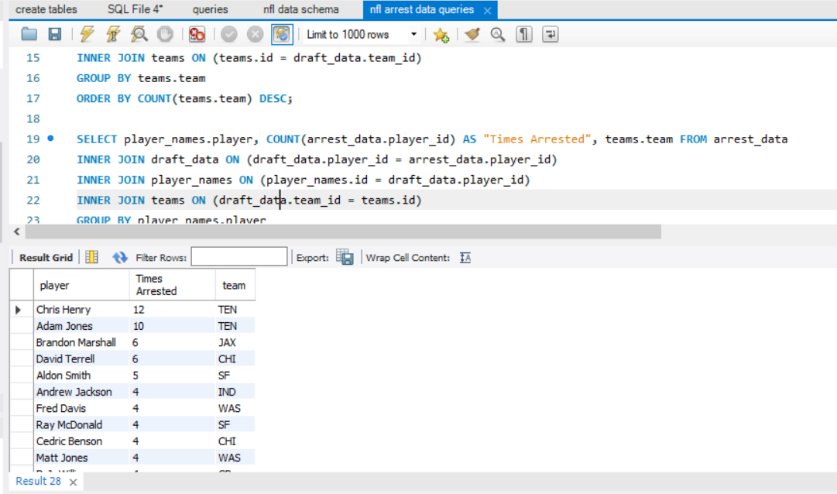
1. Players arrested by school attended:



1. Players arrested by team who drafted them:



1. Which players have the most arrests?



1. Wow, this Chris Henry seems like a mess, let’s see what he was arrested for:

