**ETL Project – Professional Sports**

I**NTRODUCTION**

Extract, Transform, and Load (ETL) dates back to the 1970’s when organizations began using multiple data repositories, or databases, to store different types of business information. As a result, the need to integrate data that was spread across these databases grew quickly. ETL became the standard method to perform this process. ETL is a type of data integration process that refers to the three steps (extract, transform and load) used to blend data from multiple sources. This process involves: 1.) extracting or taking the data from the respective sources, 2.) transforming or cleaning up the extracted data into a format that can be analyzed and 3.) load the transformed data into the target source or destination which can also be a new database.

**BACKGROUND**

After a brief brainstorming session the team of Yan, Van, Sujita, Zeinab and Escoffrey decided on doing a sports related topic for the ETL group project. The team performed a search on Kaggle and Data.world websites to determine the nature and type of sports data available. Based on the search results, the team zeroed in on using the National Football League (NFL) and National Basketball League (NBA) Historical Draft Data along with the listing of all past National Collegiate Athletic Association (NCAA) College Football and College Basketball National Championship winning teams.

The objective of the project is to use ETL processes on the respective data collected to create a new database where various queries and analytics can be performed to determine if the college an athlete attends increase the probability of being drafted and if there is a correlation between the number of championships won by a college/university and the number of players drafted.

**METHODS**

To perform the ETL class project, the following sources and tools were used:

1. Kaggle, Data.World and the NCAA websites were used to obtain the source data
2. Pandas Library for manipulation and analysis of the data;
3. Jupyter Notebook (IDE) to write the codes and for data visualizations
4. SQLAlchemy an object relational mapping tool in Python used to map the data to MySQL MySQL a relational database used to store the data and from where queries are ran.
5. Python program was used to perform (run) the overall project
6. Microsoft PowerPoint a presentation program to develop the presentation slides.
7. Github a web based hosting service used as the repository

***Step 1) Extraction***

The following data files were obtained on the dates indicated from the respective websites:

* May 1, 2019: <https://data.world/aaweiss/nfl-draft-history>
* May 1, 2019: https://www.kaggle.com/pmp5kh/nba-draft-19802017 NBA Draft 1980 - current
* May 1, 2019: https://www.ncaa.com/news/basketball-men/article/2019-04-08/ncaa-mens-basketball-champions-1939-today NCAAB
* May 4, 2019: <https://www.ncaa.com/history/football/fbs>

The data was either downloaded as a csv file or converted to csv files and used as the source data files.

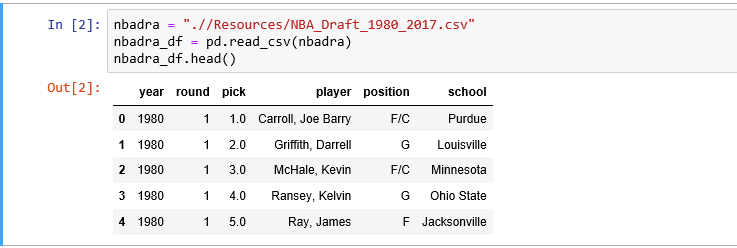
***Step 2) Transformation***

The source data as extracted are not usable in its original form and as such need to be cleansed, mapped and transformed. The transformation of the data was done in two parts. Part I was done in the respective csv files and Part II done using pandas. To cleanse the data the following data integrity procedures were performed:

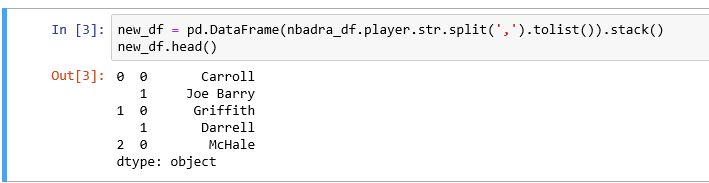
Part 1: These limited review and validation were performed for each of the csv source file. Identified and removed special or foreign charters that were contained in the players’ names, school names and coaches’ names, duplicate and fragmented data. Checked and ensured that the naming convention of the colleges were consisted on each file. Performed spell checks and corrected any miss spellings. Ensured that all dates were in the same format and consistent on each file. Information for years prior to 1980 was deleted from files that contained such data.

Part II: The updated csv files were read into Pandas using the Jupyter Notebook as separate dataframes and the following manipulation procedures were performed (Note , results of the NBA data transformation are included as examples):

Read in files and view headers



Data type check



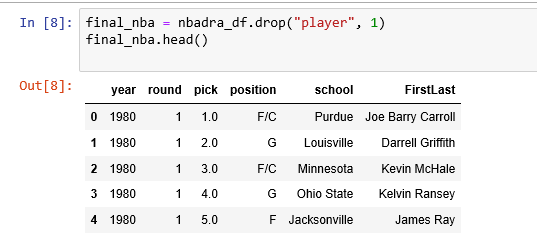
Split a column into multiples, create new dataframe, add column and concatenate columns



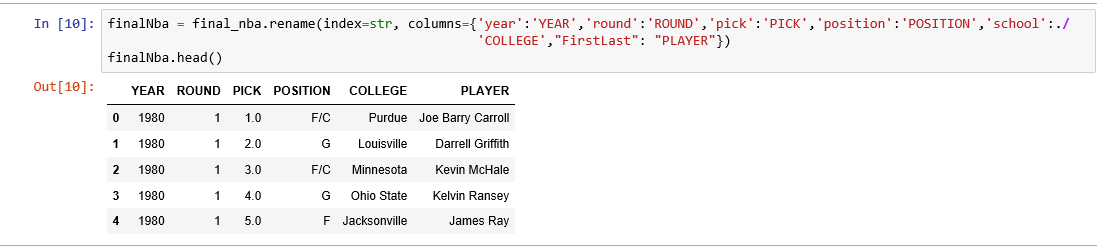
Merge two dataframes



Drop columns

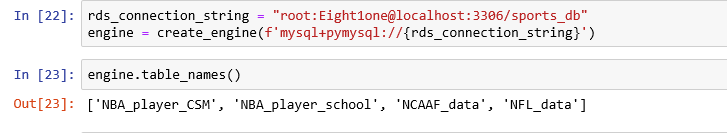


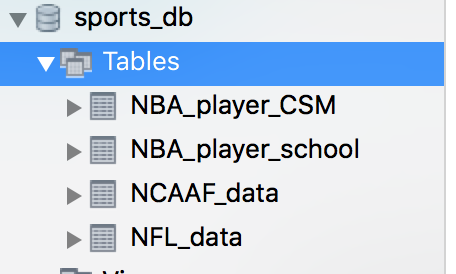
Rename and format column headings



***Step 3) Loading***

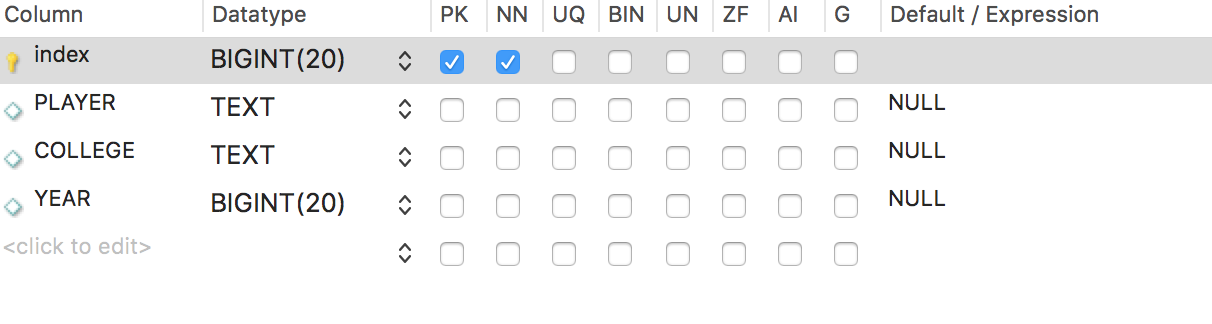
Load data and create new database with four tables





**Set Primary Key:**

In order to set a Primary Key, we had to create and index and then assigned the primary key to that index.



**QUESTIONS, QUERIES AND QUERY RESULTS**

To test the functionality of the Sports databased created, the team can up with several questions, wrote queries to the database to generate the results listed.

**Question 1:** Which Colleges have the most players drafted by the NBA between 1980 and 2018?

**Query 1:**

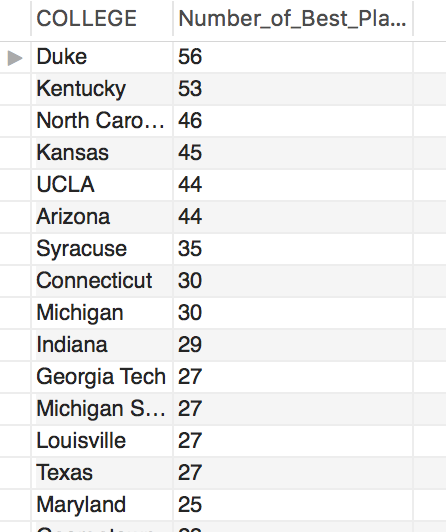
*SELECT COLLEGE, COUNT(\*)*

*FROM nba\_player\_school*

*GROUP BY COLLEGE*

*ORDER BY COUNT(\*) DESC ;*

**Result of Query 1:**



**Question 2:** Which are the top Colleges for players drafted by the NFL between 1980 and 2018?

**Query 2:**

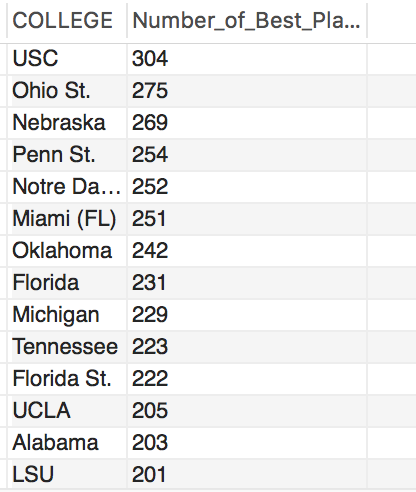
*SELECT COLLEGE, COUNT(\*)*

*FROM NFL\_data*

*GROUP BY COLLEGE*

*ORDER BY COUNT(\*) DESC ;*

**Result of Query 2:**



**Question 3:** What are the names of the players drafted by the NFL from the 2016 Championship team?

**Query 3:**

*select NCAAF.YEAR, NCAAF.CHAMPION,NFL.Player*

*from NCAAF\_data NCAAF*

*LEFT join NFL\_data NFL on (NCAAF.CHAMPION = NFL.COLLEGE )*

*WHERE NCAAF.index = 2 and NFL.YEAR =2016;*

**Result of Query 3**:



**CHALLENGES AND LIMITATIONS**

The following are limitations regarding the Sports ETL class project:

* Time and experience
* The source datasets obtained and used were limited to only the two most popular men sports in the U.S. namely Basketball and American Football.
* The date range was limited to 1980 to 2018.
* Quality of the source data and the time needed to cleanse.
* Inconsistencies in the data type, style and format as a result of the different sources from which the data was obtained.
* Setting the Primary Key: As there was no unique identifier that was common to each dataset, the team had to set an index to use as the primary key.
* Not players were drafted from college (higher school and international)
* Spelling of names, foreign characters included in names and the inconsistencies of the college names (Texas, UT, University of Texas).

**OPPORTUNITIES**

With time and more experience the team could expand the source data to include Women College and Professional sports such as Basketball and Soccer and as such create a more comprehensive database. Also, there is the opportunity to run more queries including more advance queries and to perform various analytics. Data visualization could be used to enhance the presentation of the results. Other tools and programs could also be utilized.

**CONCLUSION**

Based on the ETL procedures performed, the Sports data based created, the queries ran and the results returned; we are able to conclude that the Sports database is functional. USC is the top rank college for players drafted by the NFL. Duke is the top rank college for players drafted to the NBA. There were 304 players drafted by the NFL from USC and 56 players drafted by the MBA from Duke during the period 1980 to 2018. There were 11 players from the Clemson 2016 National Championship Team drafted by the NFL.