ETL Project – Professional Sports

INTRODUCTION

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 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.

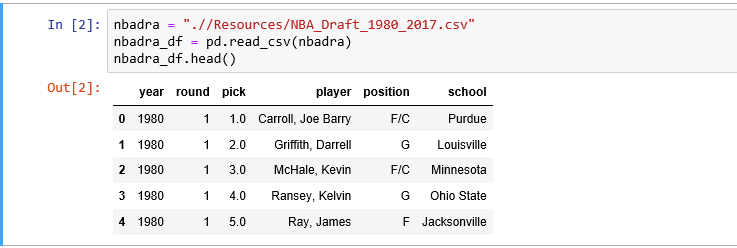
Step 2) Transformation

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:

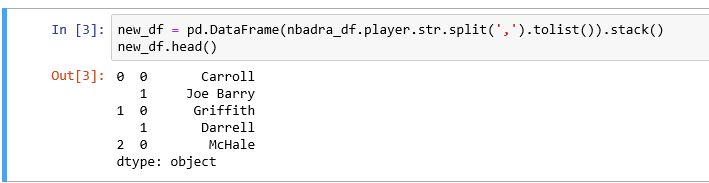
Part 1: These limited review and validation were performed for each of the csv source file. Identified and removed special or foreign charters that were contained in the players’ names, school names and coaches’ names, duplicate and fragmented data. Checked and ensured that the naming convention of the 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



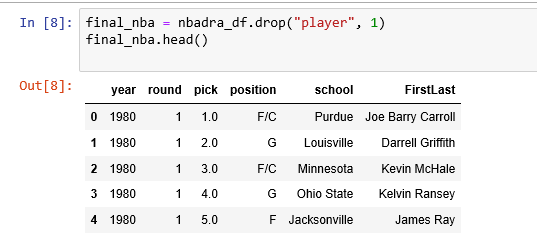
Split a column into multiples, create new dataframe, add column and concatenate columns



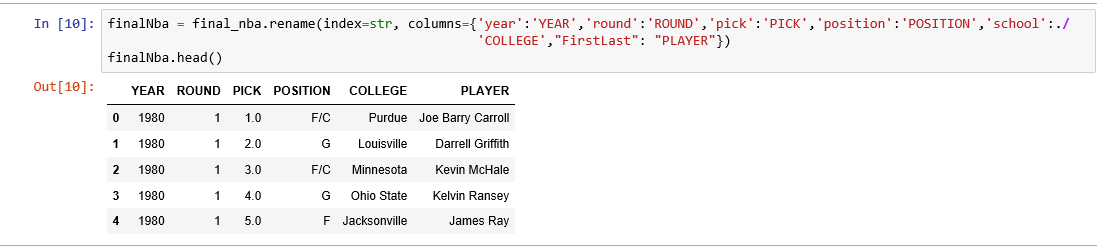
Merge two dataframes



Drop columns

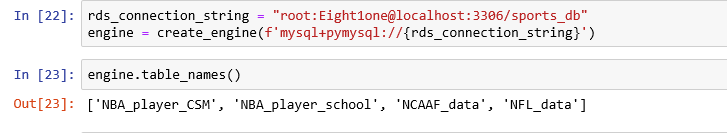


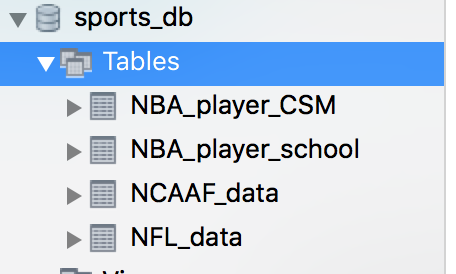
Rename and format column headings



Step 3) Loading

Load data and create new database with four tables





Set Primary Key

QUESTIONS/QUERIES

RESULTS AND OBSERVATIONS

LIMITATIONS

The following are limitations regarding the Sports ETL class project:

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 2017. Inconsistencies in the data type, style and format as a result of the different sources from which the data was obtained.

OPPORTUNITIES

CONCLUSION

To be deleted

1) Description of table keys (primary/foreign)

2) Limitations. Every analysis, dataset, and model has limitations. Get used to acknowledging them.

3) Verbose explanations describing functions used in queries

4) Screenshots of results. Prove to me your code works

5) Conclusions/Future Work. Need this on every report. Again, no analysis, dataset, or model is the final word on anything. Future work can always be done, so acknowledge that

Validations are done during this stage

* Filtering – Select only certain columns to load
* Using rules and lookup tables for Data standardization
* Character Set Conversion and encoding handling
* Conversion of Units of Measurements like Date Time Conversion
* Data threshold validation check. For example, age cannot be more than two digits.
* Data flow validation from the staging area to the intermediate tables.
* Required fields should not be left blank.
* Cleaning (for example, mapping NULL to 0 or Gender Male to "M" and Female to "F" etc.)
* Split a column into multiples and merging multiple columns into a single column.
* Transposing rows and columns,
* Use lookups to merge data

The new dataframe files created were also saved as csv files to be used in the plotting graphs and charts to showcase the relationship between the respective variables being analyzed.

Load verification

Ensure that the key field data is neither missing nor null.

Test modeling views based on the target tables.

Check that combined values and calculated measures.

Data checks in dimension table as well as history table.