**Technical report for ETL-Project**

**NBA Data / Team 4**

**William Clewett, Pushpraj (Raj) Rana and Paul Murray**

**Data Sources:**

For our project we used data about the NBA from the following sites.

[www.kaggle.com](http://www.kaggle.com)

[www.basketball-reference.com](http://www.basketball-reference.com)

<https://data.world/>

We extracted data from each site as .xlsx and .csv files. And data scraped information from Kaggle

We searched other data sources but found what we needed from the sites listed.

**Importing the data**

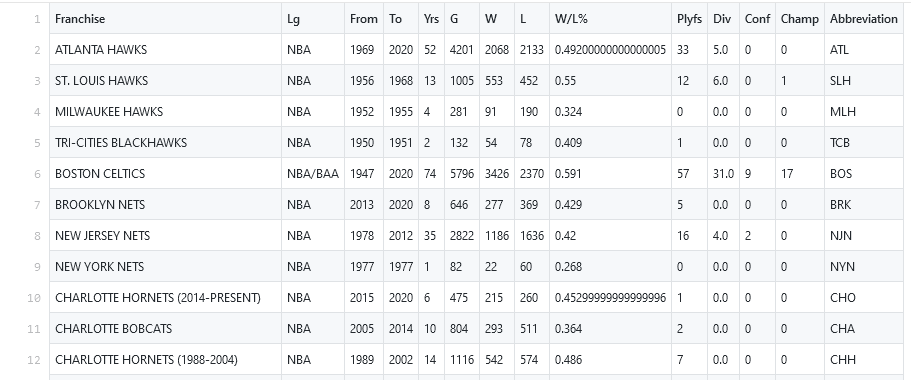
We created the database and tables to support the import of the data. Some of the data was missing and date ranges were variable and there were blank rows. We created the database and tables to support the import of the data into Postgres. There were some issues at the end of the Postgres process, and we had to make changes in the Pandas/Jupyter tables to finish the import process consolidated it for the final Postgres import process.

**The type of transformation needed for this data (cleaning, joining, filtering, aggregating).**

The original files pulled and scraped were converted to .csv files.

Some were .xlsx files and the scraped files were taken into excel and saved as .csv files

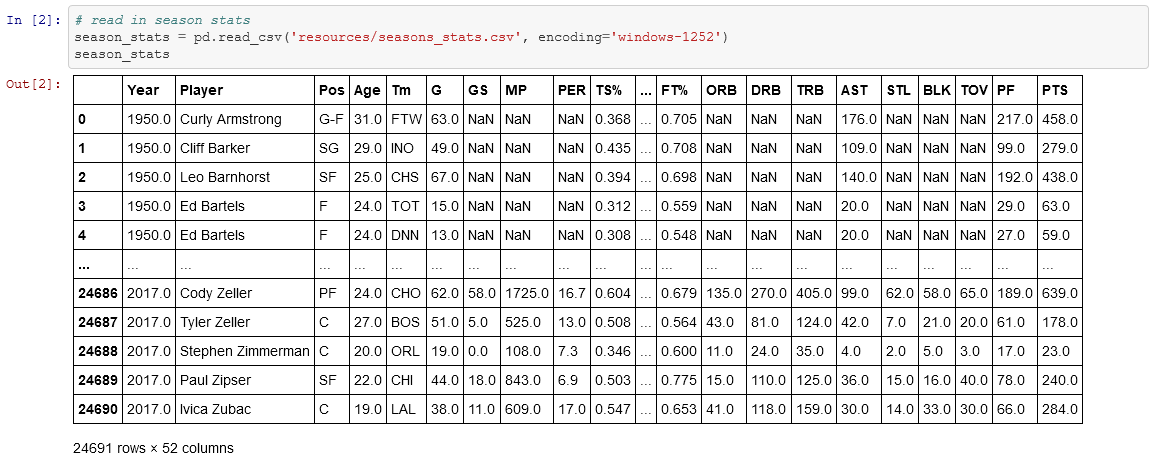
We cleaned up the .csv files by removing un-needed columns and rows.



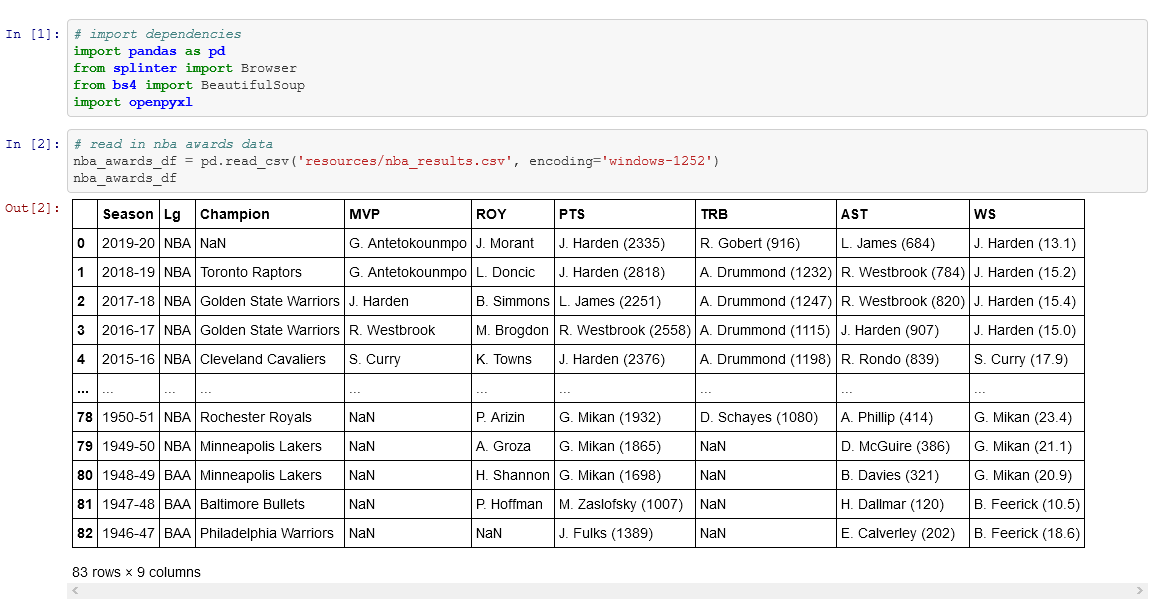
Created unique player\_ID’s based on the College they attended.



We used Pandas to analyze data types for counting stats.



We used Pandas and Beautifulsoup to read in the awards data

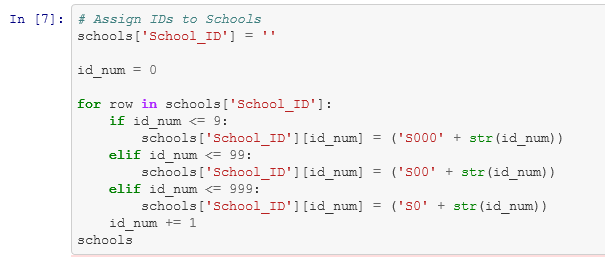


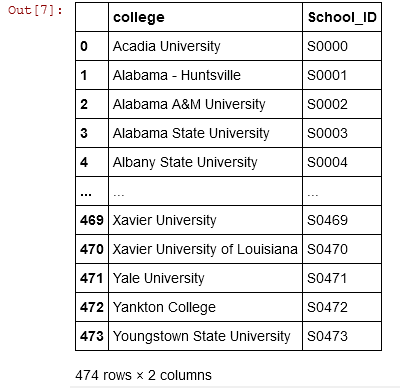
We joined tables so each players NBA team and College could be cross referenced



Each row corresponds to team and contains information on the player and the College they attended.

We also created a unique ID for each College.





**The type of final production database to load the data into (relational or non-relational).**

**We created a relational database.**

We loaded into Postgres the database\_files and the transition\_files.

**The final tables or collections that will be used in the production database.**

Postgres to load both the NBA and NCAA data and allow it to be searched and compared

We used:

Postgres, Jupyter Notebook and Python packages for our work:

Pandas - For data manipulation, reshaping, merging, sorting, etc.

Splinter/ Beautifulsoup - Join, query and update the database

Postgres - For loading the data into the database and running queries against the same.

SQL – For table creation