Week 13: ETL Project

Group 1: Randall Teves, Ian Koenes, Christina Coder and Ricardo Cano

For our ETL Project, our group sought to prepare a database that could be used to analyze potential correlations between Division I college basketball teams that advanced to the final rounds of the NCAA and the NBA draft picks those teams produced for their tournament runs’ corresponding year.

In order to delve into this possible analysis, our group used Google Dataset Search (https://datasetsearch.research.google.com/) to find two historical datasets that were originally in csv format. One dataset documented NCAA tournament results for every tournament game between 1985 and 2019; the other dataset listed every NBA draft pick and their college career statistics between 1976 and 2015.

Our group imported the datasets into VSCode and used pandas to transform the data. After creating data frames for both sets of data, our group then began cleaning the data. Both sets of data included many columns of information not necessary for analysis (the draft pick file included 35 columns of data). So, we cleaned the data partly by using the .loc function to remove the unnecessary columns.

After cleaning the data, our group created a connection string linking our file in VSCode with the SQL file we created in pgAdmin4, making sure to check that the connection was made successfully. In pgAdmin, we created two tables – “tournament\_games” and “draft\_picks” – to house the two sets of data that we cleaned using pandas.

After we loaded our data into pgAdmin, our group then joined the tables together to create a database that included information on draft picks and NCAA tournament outcomes. The main.ipynb file in our submitted repository includes the code used to clean and transform the original data. The query.sql file includes the queries we ran in pgAdmin to load the data and join the datasets together.

One of the most difficult challenges in this effort was the fact that the two datasets we used did not share a relational data point that we could use to reliably join the tables. There was no field of data we could use as a primary key. Though the datasets both included information on players’ colleges, it was not documented in a consistent manner. For example, whereas one set of data listed a player’s college as the University of Washington, the other dataset identified that same university as “Washington.”

Because the two datasets we used came from two different sources – and because they did not share a relational field – there are potential caveats that should be considered for analysis. For example, it is difficult to reliably draw conclusions whether individual players’ performances in the NCAA Tournament affected their NBA Draft result because the original tournament data did not include statistics on teams’ or players’ performances -- or whether they were even active -- in the Big Dance beyond the final scores.