Jeff Lyng

Irene Zou

Bradley Higdon

Sami Abu-Douleh

John Marriott

Data Analytics Boot Camp

28 October 2020

Hockey Metrics - ETL Project (Group G)

Extraction

We found 20 files containing a total of 290 columns of data on a Kaggle page regarding hockey. Subgroups of these files focused on elements such as scoring, players, teams, coaches, and other topics. In this project, we used the data in the Master.csv file from Kaggle. For our second source, we found around 7,200 players and career stats on Elite Prospects which we scraped directly from the website into a CSV file we created.

Transform

The goal was to align the Kaggle CSV with the Elite prospects CSV by player name (unique field). In order to align the data, we had to remove special characters (ex: ö, É, etc.) and positions from Elite Prospects Hockey Stats names. We also separated the period column to filter out players by first year and to match with the other data sets.

For the data from Elite Prospects, we created career\_stats\_adj3.cvs by combining the 73 pages data from the URL. While combining, we encountered some challenges because this particular data set has column names that are different from the majority of other data sets we had ready.

That includes missing values which caused errors in the code; the Name and POS were combined using “( )” in this set, special characters were in the Name column; the period column is a range instead of first and last year; and the mismatch data from years prior to 2011.

In order to create our database with this data set in conjunction with our main source “master.csv”, we had to put a lot of work in to drop the missing values, split the Name and POS into two columns, replace the special characters from the Name column, and split period column into “firstYear” and “last Year”. Additionally, we filtered out players entering the league after 2011 to match the master file and remove the duplicate names.

For master.cvs, it has some different arrangement for its columns that caused some errors while we were coding as well. It has the first and last name separated, which is not matching with the career\_stats\_adj3.csv. To solve this and to join the data on the name values, we concatenated the firstName and LastName columns into one Name column and reordered columns. To match the players’ name column, we had to filter out any player that did not have a firsNHL year and filtered out all goaltenders since that position was not included in the stats CSV. As well as removed the duplicated names.

Load

First, we created a diagram to give an outline of what our database would look like in SQL. We had to make some minor edits to the code that the diagram on quickdatabasediagrams.com gave us when we exported it in order to make the joins work properly, but otherwise the diagram fits what our database looks like. Our final database contains two tables; a Master\_adj table to contain the basic information of players such as first\_name, playerID, and POS, while the career\_stats\_adj3 table contains the more detailed statistical information such as GP, G, and A. These two tables were created by importing the two adjusted CSV files into SQL.