ETL Project

August 2019

Yang Liu, Justin Roth, Bill Schreiber, Tolly Smith

For our project, we chose to look at NHL data available through Kaggle:

<https://www.kaggle.com/martinellis/nhl-game-data>

There are 10 sources of data to consider. All of them are in csv format.

Through examination we considered which fields are included with each source and selected two files for this project: game.csv and team\_info.csv

Next, we **EXTRACTED** these files from the Kaggle site.

Game.csv has 16 data fields:

|  |
| --- |
| game\_id |
| season |
| type |
| date\_time |
| date\_time\_GMT |
| away\_team\_id |
| home\_team\_id |
| away\_goals |
| home\_goals |
| outcome |
| home\_rink\_side\_start |
| venue |
| venue\_link |
| venue\_time\_zone\_id |
| venue\_time\_zone\_offset |
| venue\_time\_zone\_tz |

Team\_info.cvs has 6 data fields:

|  |
| --- |
| team\_id |
| franchiseId |
| shortName |
| teamName |
| abbreviation |
| link |

The data included more data than we need for our queries. We then set out to **TRANSFORM** the data using Jupyter notebook by dropping unnecessary fields. We also needed to change all of the fieldnames to lowercase for use in pgAdmin. The fields we kept are listed below:

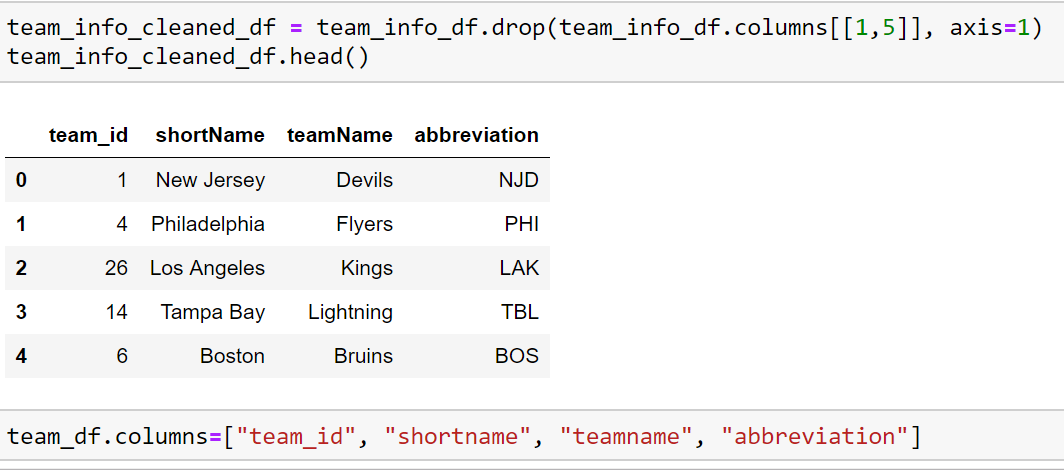
In game.csv, 8 data fields were retained:

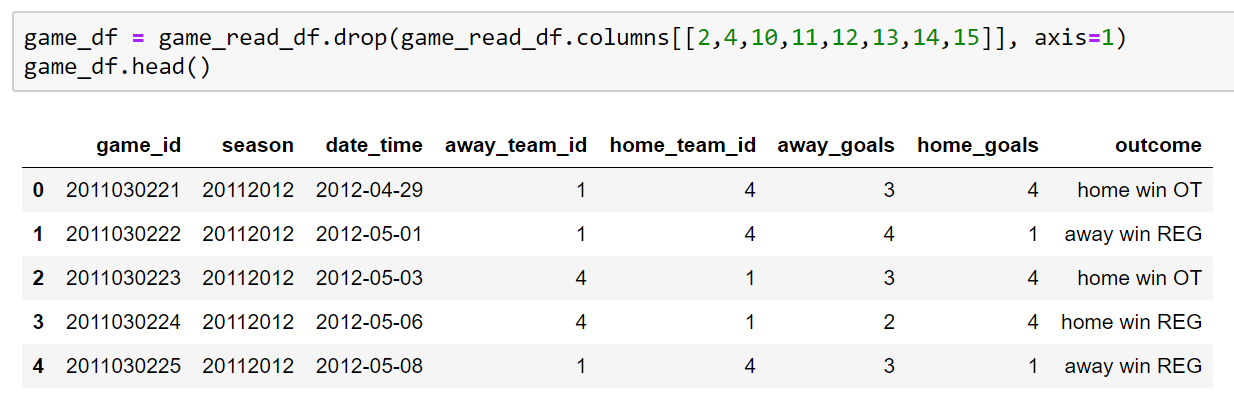
|  |
| --- |
| game\_id |
| season |
| date\_time |
| away\_team\_id |
| home\_team\_id |
| away\_goals |
| home\_goals |
| outcome |

In team\_info.cvs, 4 data fields were retained:

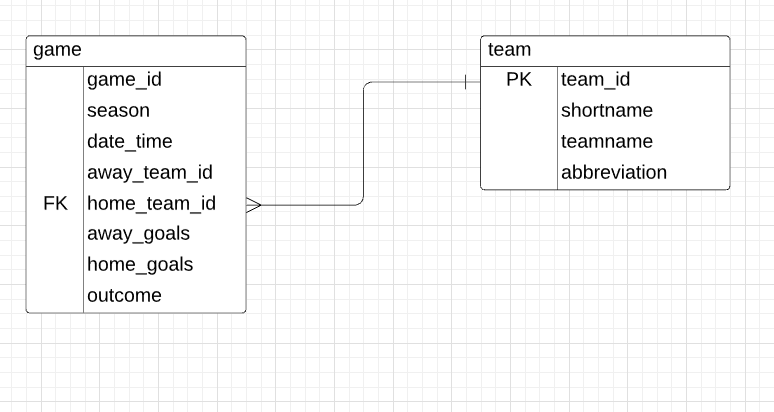
|  |
| --- |
| team\_id |
| shortName |
| teamName |
| abbreviation |

Some of the python code used includes:





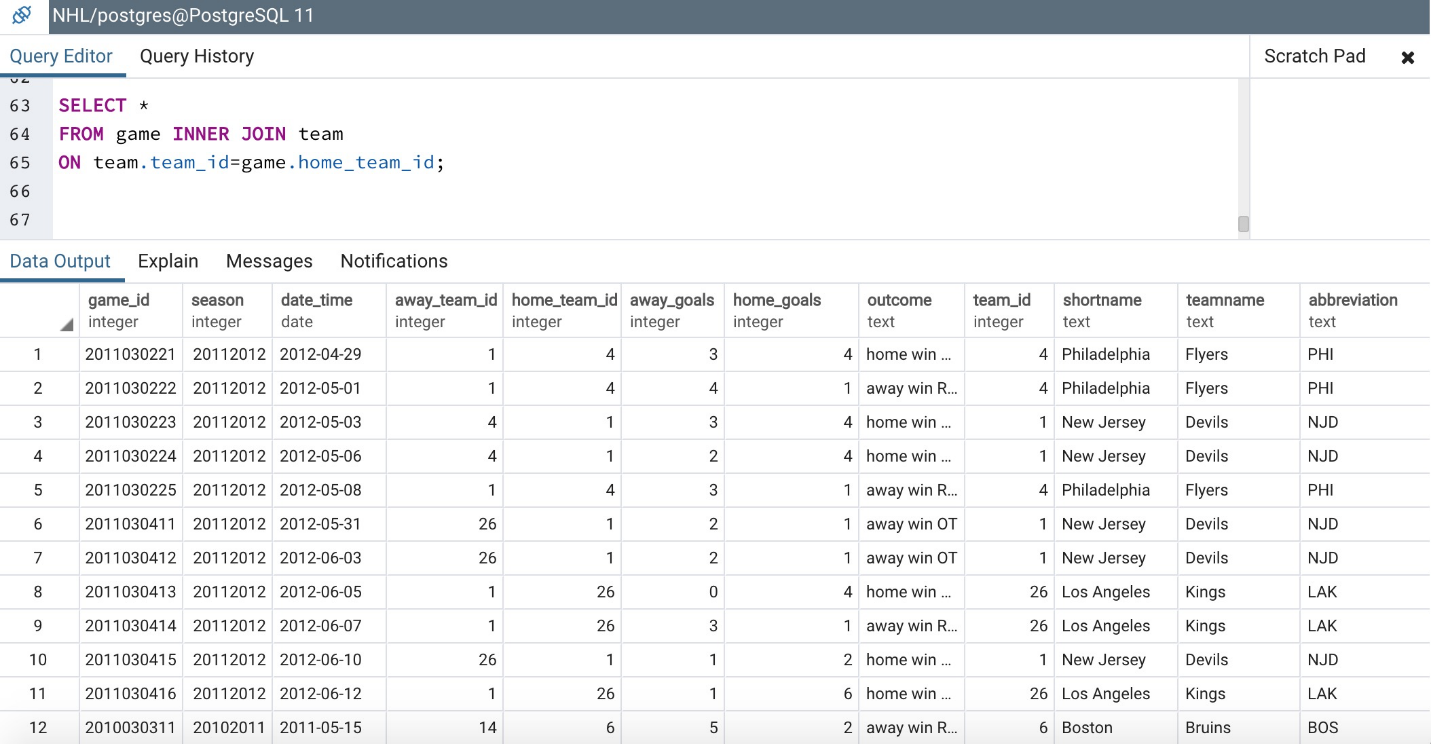
We then created an ERD in Lucid Charts to identify primary keys and foreign keys in the relational tables.



The transformed files were saved as a csv from Jupyter notebook and imported into pgAdmin for use in our queries. The files were created in pgAdmin and joined into one table. This resulting table is the **LOAD** portion of our ETL project.

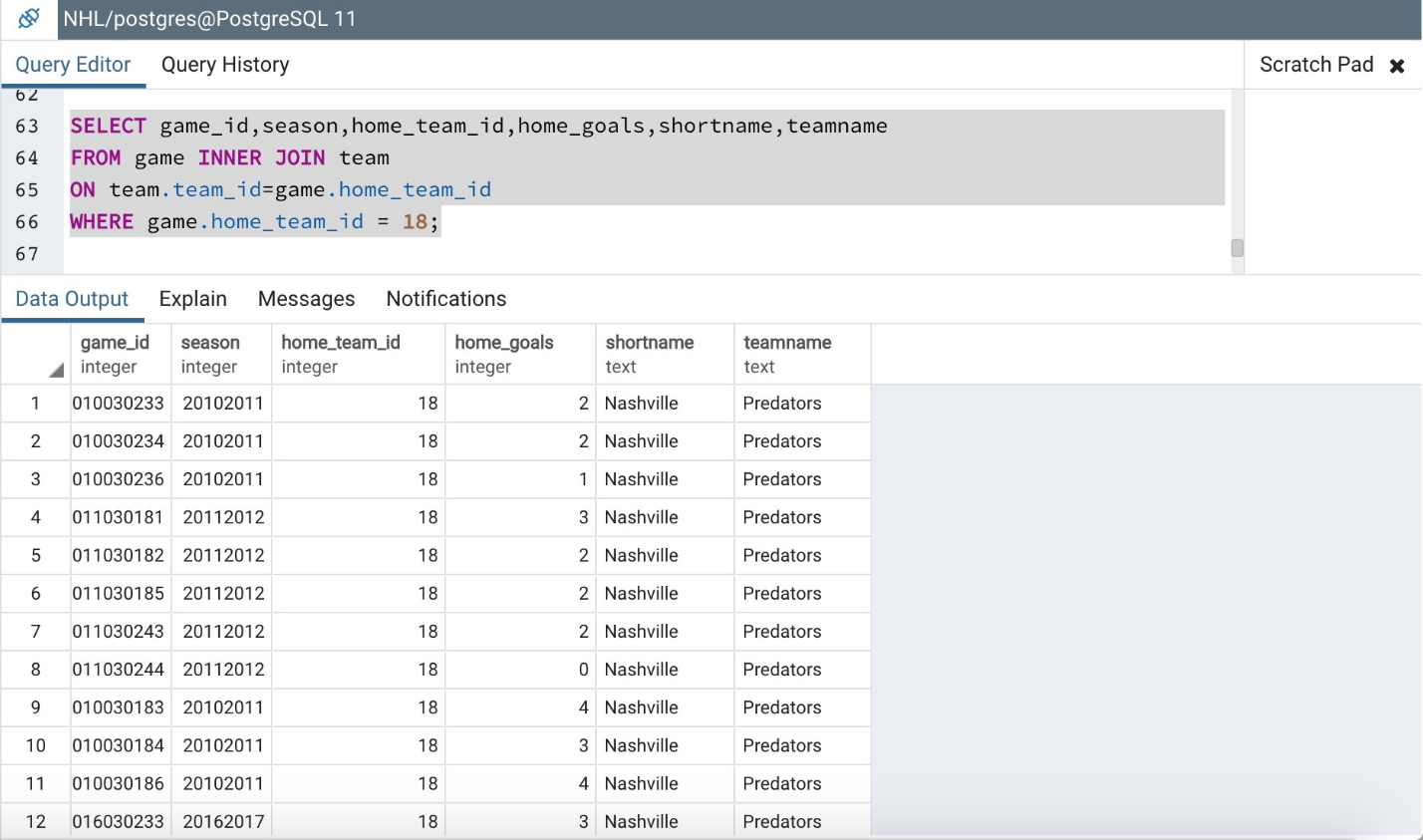
This project enables queries on NHL data at the team level. Incorporating the team.csv file provides critical information so we can identify each team by something other than a randomly assigned identification number.

Code and result from joining tables:

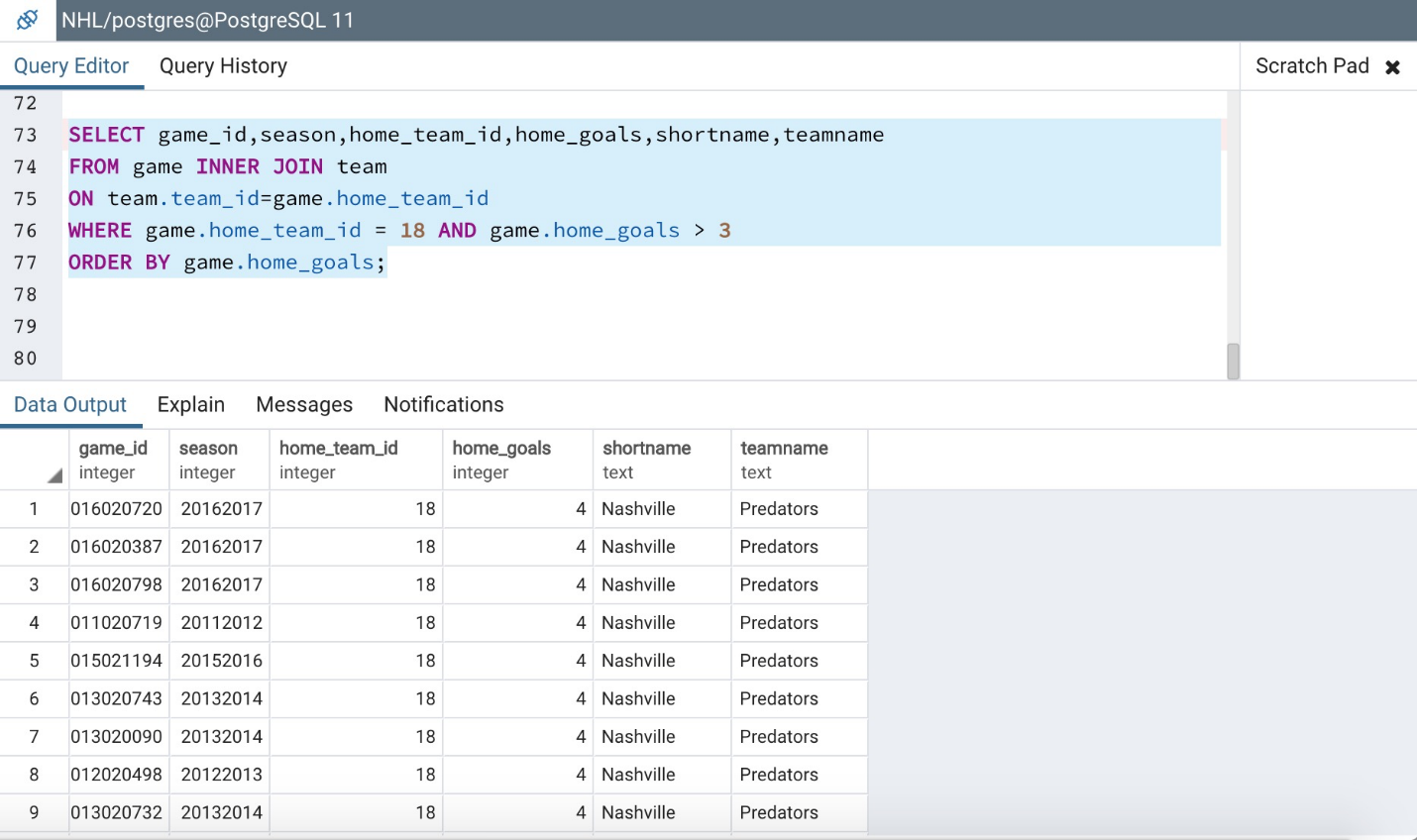


Queries:

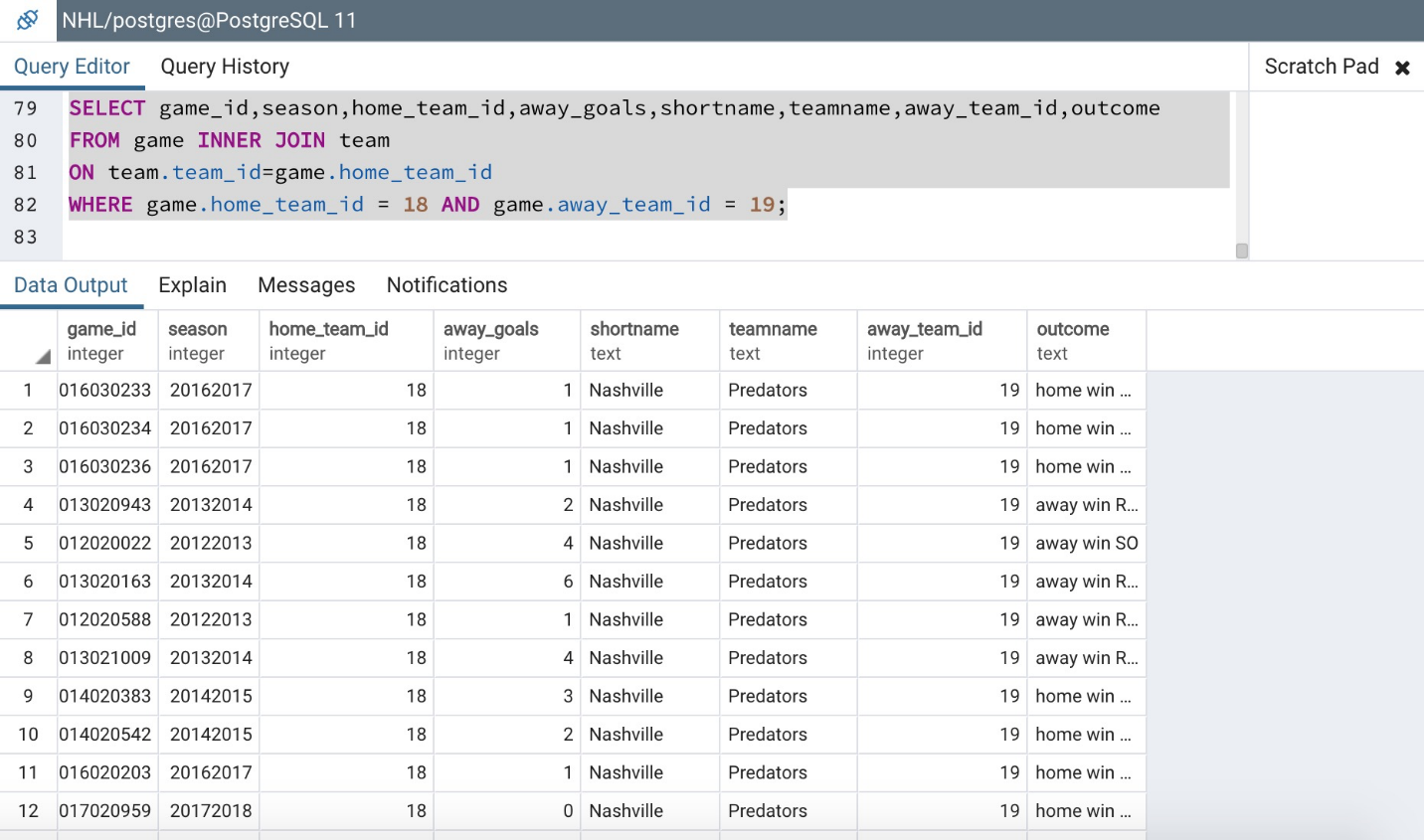
1. All Nashville Predators home games



1. Nashville home games when they scored 4 or more goals



1. Nashville home games versus St Louis



1. Nashville home wins in regulation

