ETL Project

Overview

Using ETL we wanted to create a database that contained Season stats and attendance record for the 2011 NHL season.

1. Extract
   1. For the Extract portion of this project we extracted 2 data sets. The first dataset was extracted from dataworld.com containing the attendance records for NHL teams from 2001 to 2018 seasons. The second dataset was extracted from Kaggle.com that contained season stats from each NHL team spanning from 1909 to 2011.
   2. Data set links;
      1. <https://data.world/makeovermonday/2019w1>
      2. <https://www.kaggle.com/open-source-sports/professional-hockey-database#Teams.csv>
2. Transform
   1. For the Teams.csv we first grabbed the columns of the dataset we wanted. We selected the columns containing year, rank, name, ConfID, and DivID. Then we performed a “groupby” to select only the data from 2011 season. Next we renamed columns to clean up our naming system so it will cohesive across both datasets. Finally we performed a slipt function on the team names to create two separate columns, one containing the city and the other containing the team name.
   2. For the NHL\_Attendance.csv we started selecting the columns Season, Rank, Team, Home Attendance, Road Attendance, and total Attendance. Next we renamed the columns to have cohesion across the datasets.
3. Load
   1. Created a database within Postgres then made a connection variable in pandas. Then we imported the csv files and converted them in to a database using the panda query “to\_sql”. Finally using a “.read\_sql\_query” we confirmed two tables were created and saved into Postgres. We performed a join on the two tables and postgress kept on giving us an error. After we went through a changed all the column titles to lowercase letter the join within postgres worked perfectly.