John Allison

Andrew Shaw

Yuji Tang

ETL Project Report

Extract

* We chose a dataset of PGA tour gold statistics from 2010 to August 2019, which we obtained from Kaggle. The tables were originally pulled in a CSV file format.

Transform

* For the transformation, we pulled the CSV files into Pandas. The 2019 data table had a column labeled ‘Season’ while the historical table included a ‘Year’ column. To match the tables, we pulled the year from the season column in the 2019 table using a Datetime function. The ‘Value column in both of the tables included several characters which were giving errors. To filter these values, we used Regex.

Load

* After we cleaned the data sets, we imported the data frames into Pgadmin where we joined them into one database.