**Project Report - ETL**

* **Extraction**

The extraction process for the project included downloading .CSV files from two sources; Kaggle and DataGolf.com.

Using Kaggle I found an immense about of golf data for the 2018-2019 season. The CSV file included over four thousand individual metrics for each golfer who is consider a professional (roughly 200 players).

Using DataGolf.com I obtain a CSV file to extract metrics/stats for professional golfers from the 2017-2018 season. This file was much smaller with roughly 30 different stats and a much more friendly format. Each stat was its own column as oppose to the previous file which had all the stats in one column.

* **Transform**

The 2018-2019 CSV created the biggest challenge when cleaning. Rows needed to be converted into columns, duplicated needed to be removed and NAN needed to be dropped. To do this I created a data frame in Pandas and kept only the metrics I wanted to combine with the 2017-2018 dataset. I was anticipating joining the two data sets and was unsure which Primary Key I would want to use so I renamed all the columns so I could decide later.

My goal for the transformation portion was to have two data frames (and new CSV files) with identical column names for the 6 metrics I was interested in.

Once the transformation was complete, I could load the data frames into a database tables (using pgAdmin) and join them.

* **Load**

Loading the data frames was tricky. I struggled with converting them to SQL. I chose pgAdmin because it seemed to be the easiest to use and I thought using a structured DB made sense. I created a table for each dataset which was pretty straight forward.