ETL Project

**Bryant Springer** and **Lisa Godfrey**

Data Sources: https:\\data.world

We decided to search for tennis information and came across two .csv files that we could use to tell an informative story. We wanted to determine who won the most Grand Slam Championships out of each gender. One of the files provided information about the results of the Men’s Grand Slam Championships and the other file provided information about the results of the Women’s Grand Slam Championships.

**Extractions**

**Files that we extracted from:**

grand\_slam\_championships\_champion\_vs\_runner\_up\_mens\_singles\_1968\_2018.xlsx

grand\_slam\_championships\_champion\_vs\_runner\_up\_womens\_singles\_1968\_2018.xlsx

Once we decided on which files we wanted to use, we stored each of the two files into two separate excel files as .csv files. The .csv files were named ‘Womens\_GrandSlam.csv’ and ‘Mens\_GrandSlam.csv’.

We then stored each of these files into separate data frames, using pandas. The data frames were named ‘Womens\_df ‘ and ‘Mens\_df’.

**Data Cleanup and Analysis**

**Type of transformations needed for this data:**

Next, we created two new data frames, each with select columns from the original two data frames listed above using the code below.

womens\_grand\_slam\_df = womens\_df[['Year', 'Major', 'Champion']]

mens\_grand\_slam\_df.rename(columns={'Major#': 'Major'}, inplace=True)

As you can see above, a bit of **cleaning** needed to be done to the mens data. The ‘Major#’ column needed to be renamed to ‘Major’. So we chose to make that change in the data frame before loading it into the database.

We then merged the two data frames above by performing an outer join on Major and Year to create another data frame called ‘grand\_slam\_df’ which contained the Year, Major, Mens champion and Womens Champion using the code below:

Mens\_Womens\_Champion\_df = mens\_grand\_slam\_df.merge(womens\_grand\_slam\_df, how='outer', on=['Year', 'Major'], suffixes=('\_Mens', '\_Womens'))

We then loaded this info as a table into the database and called it ‘GrandSlamChamps’.

Mens\_Womens\_Champion\_df.to\_sql(name = 'GrandSlamChamps', con = engine, if\_exists = 'replace', index=False)

Next, we decided to show the names of all women champions and the total number of championships each of them has won. This required us to use some aggregation to count the number of wins for each of the women and men. We used this information to plot a bar graph and repeated this for the Men’s data as well.

We chose to use a **relational database** because our data was in a **structured** format with similar fields and column names and types. We created a database called ‘Tennis’ that included three

tables called ‘Mens’, ‘Womens’, ‘GrandSlamChamps’ were created using SQLAlchemy. Our data transformations were performed using panda data frames prior to loading the three tables into the database. It was easier transforming our data using Pandas and then loading afterwards! Based on our data, we determined that Roger Federer won the most Men Grand Slam Championships. And Serena Williams won the most Womens Grand Slam Championships.