Jonathan Rivera
Ed Torrente
UCI Data Analytics

ETL Final Report

For our second project, our team decided to work with NBA stats in order to compare two of the games biggest stars: Lebron James and Steph Curry. In this case, we decided to look at playoff data specifically, and dig into certain scoring metrics that are measured when looking at offensive productivity. Ed found a CSV file with Lebron's playoff statistics, while Jonathan found a Kaggle dataset containing the CSV for Steph's playoff performance broken down game by game throughout his career. Both datasets were imported and cleaned using Pandas in our Jupyter Notebooks. Ed's dataset working with Lebron's playoff data contained a large amount of null values, so he had more initial cleaning to do in order to get the correct dataframe desired for our project's necessary statistics.

After working through our notebooks, we were able to extract the columns that would be our tools of comparison: points scored, three point percentage, and playoff career high for a single game, found using the max() function with our scoring column. Lebron's playoff career high was 51 points, while Steph's career high was 47. We then took the average function to find average scoring total for each players playoff career, with Steph averaging 28.9 PPG while Steph averages 26.5. Once our jupyter notebooks had similar columns to merge with, we used SQL to join both of our tables using Postgres. We first had to use grouping in order to group playoff games according to the season played. This way, we were able to join under the same column, "Season_year". We also used the order by function in SQL to get a clearer picture of the merged table containing both Lebron's and Steph's stats side by side.

Finally, after working through Pandas and SQL, combining data into a single database, we went back into our notebooks to connect to our SQL databases, in order to run queries directly from our code. We created our necessary connections and engines, and ran various queries to get the data from our SQL database. While we experienced our biggest challenge using SQL joins, eventually we were able to smoothly extract, transform and load our data.