January 25, 2021

GROUP 7 PROPOSAL:

Participants: @Alvin McCray @JB Kinlacheeny

**DATA SETS:**

NFL Statistics Scrape data from Kaggle: https://www.kaggle.com/kendallgillies/nflstatistics?select=Basic\_Stats.csv

And NBA Players by State data from Kaggle: https://www.kaggle.com/rishidamarla/nba-players-by-us-state

Update: 1/27: We realized we cannot merge the original datasets and only used the NBA datasets.

New datasets:

1. NBA Players stats since 1950. <https://www.kaggle.com/drgilermo/nba-players-stats?select=Players.csv>
2. Social Power NBA. <https://www.kaggle.com/noahgift/social-power-nba?select=nba_2017_twitter_players.csv>
3. NBA 2017 Salary. <https://www.kaggle.com/noahgift/social-power-nba?select=nba_2017_salary.csv>

**EXTRACT:**

Both files will be downloaded as CSVs (Alvin: 10 min)

**CHALLENGE:** Alvin will load one of the CSVs into Postgres using PG Admin’s import tool (Alvin: 15 min)

One CSV will be imported to pandas using from\_csv, the other using from\_sql with pandas (JB: 25 min)

Update: 1/27 – We both imported all the CSVs into Pandas. We imported all 3 datasets above. We transformed the datasets by using the rename function of the Player/NAME column to “Player.” We then created Pandas DateFrames for all 3 CSVs and only used selected columns. We dropped null values using the dropna function. We tried to merge all three dfs on=”Player” in one pd.merge function and got an error. We then merged two CSVs and were able to merge on=”Player”. We then merged the last CSV to the merged df on=”Player” and was successful, however, the df only consists of 3 rows.

**TRANSFORM**

Once imported, we will join the tables on State/Birthplace Player\_ID, state\_id, city\_id (JB: 15 min) Update: 1/27 – we eventually joined the tables on Players.

Rename all columns per instructions (see load) (JB:10 min) 1/27 – Done in pandas.

**OPTIONAL:** We will look to clean the joined data and select data only from DC if time permits. – the original dataset was removed.

**LOAD**

While JB is extracting/loading, Alvin will read through each dataset and create a Postgres table to load into. (Alvin/JB: 1 hr)

JB will work with Alvin to rename columns appropriately (JB: 15 min)

Alvin will write the portion of the python script that loads the data into the Postgres database. (20 min)

Notes:

This is slated for about 1.5 hours of work for each of us; we plan to use extra time Wednesday to explore any bugs and verify data integrity. Project to be complete Wednesday night.

kaggle.com

NFL and NBA player birthplaces.

Kaggle is the world’s largest data science community with powerful tools and resources to help you achieve your data science goals.