*Steven Rhee, Eric Schoonmaker, Soham Patel*

**NBA Dataset from 1950 to 2018: ETL Report**

1. Sources of data

* We extracted the CSV files from Kaggle here: <https://www.kaggle.com/drgilermo/nba-players-stats>.
* Raw data includes three separate CSV files consisting of player information and season stats from **1950** to **2018**.
* Two out of the three CSVs contained overlapping information, and as a result, the one with less information was ignored for the ETL project.

2. Extract:

* We used Pandas pd.read\_csv function to load the CSV Files into dataframes.

3: Transformation:

* Based on the column headers, a database schema was created for data integrity and security using QuickDBD (see schema and ERD for details).
* We removed “NaN” rows using pandas dropna function, when applicable to all columns, on both dataframes
* Filtered dataframes were created based on the schema
* Multiple columns from the “Season\_stats” file ranging from 1950 to 2018 because they were identified as obscure NBA statistics and deemed unnecessary, and thus removed, again resulting in a filtered dataframe.
* Renamed percentage columns, such as “3Pt%” to “3Pt\_percent” so the sqlquery was in the correct file could correctly be loaded on to the SQL table
* Created College, Team and Position dataframes with unique values for SQL table for data integrity
* It must be noted that even though there were a lot of residual NaN values at the end of the analysis, the associated rows were not deleted as the NBA/ABA format has evolved significantly over the years, and some statistics recorded today, were not recorded 25-59 years ago.

4. Load:

* Creation of Database schema based on preview of column names from each of the two CSV Files
* Pandas.to\_Sql feature was used.
* Sqlalchemy to connect to Postgres database
* Relational database was chosen to maintain data integrity. See ERD below.

