ETL Project Report

-**E**xtract: The original data sources we used were all csv files and came from various places. The NBA player profile data and statistical data came from Kaggle.com. The geographic data came from inkplant.com and the NBA team latitude and longitude data came from wikipedia.

-**T**ransform: After importing the csv files cleaning of the data was required to drop null values and to reformat the data into the database configuration that we wanted. Pandas was used within a jupyter notebook. Columns were renamed to be clearer. Various merges were done with the player datasets. We wanted to use the NBA player stats from 2010-2015 only to use just data from a certain time period. Also the data set was cleaned to only include players born in the United States. Not all NBA players went to college so the data-frames were adjusted to account for that. Duplicate values were also removed. We also wanted the latitude and longitude data for the birth states of the players in order to map it on google heatmaps.

-**L**oad: Data-frames were cleaned and formatted in pandas then transferred to SQL tables in Postgres using an engine connection using SQLAlchemy. We ran into trouble loading the data into Postgres because of a naming error, we used capital letters where lower case was better and that solved the error we kept getting once it was fixed.