Our ETL project was designed to create a database that would make NBA analytics much simpler for the everyday fan. Given the wide spectrum of data usage whether it be; betting, fantasy basketball, or simply to monitor player production as a fan of a certain team, we felt it was necessary to create a database that would allow for the analysis of individual players. Being able to predict what players or teams will be good on a year to year basis is one of the hardest things to do in sports.

We decided to create a formula based on our thesis question: at what age do players enter their statistical primes? Are there any indicators to predict when a player is entering their statistical prime? We decided by intuition that a lot of factors go into this such as; coaching, playing scheme, supporting cast, development, etc. However, we also were able to induce that some of the more obvious factors are having better production overall, leading to more Player of the Week awards and floating in the upper percentile in key statistical categories such as points, rebounds, steals, assists, or blocks.

Extracting the data was perhaps the simplest task. We outlined what parameters we would need to set up in order to run our analysis and looked for data that met the criteria. Fortunately, there is a lot of NBA data available for free and while most of it is not too useful, we found the data we were looking for. Using Kaggle, the csv formatted data we needed was readily available for transformation. We did try to use API sources for data but that was extremely similar to the csv data we already we had so we opted not to use the API source available.

Our transformation process consisted of gathering the necessary data to compute our analysis in easy to read format. The data we found all consisted of CSV files which we cleaned using Jupyter Notebook. The cleaning process was quite tedious as we not only had to figure out what information was necessary, but also add z-scores for key categories in order to deem how certain players performed in certain categories compared to their peers and where they ranked in each respective category. In addition, we also had to use delimiters and string splits to clean the data and remove unnecessary noise for the data. We used two csv files, player of the week data and nba data dating back to 1950. With all the available categories and more, we decided that the best way to measure a player stepping into their prime would be defined by their best increase in production on a year to year basis. To calculate this, we added new columns calculating the z-scores for each category and found the difference between the previous year for each player. We considered the biggest increase in total z-score to be the age at which they had officially entered their statistical prime. We also cleaned the data by removing unnecessary columns from each data frame. After adding z-scores, we merged it back with the NBA data by creating unique ID for each player in each year to make the merge easier. After that, we were able to transfer it to a local MySQL database.

As we loaded our database into MySQL which was a relational database. We used 2 tables which consisted of player of the week data from 1985, and filtered our NBA statistical data and calculations for all players after the 1985 season. Our plan was to see when a player reaches a statistical prime and if in the subsequent years, they were more likely to be awarded player of the week awards in comparison to the years before hitting a statistical prime. We followed our intuition that players were likely to see more player of the week awards in the season of their best statistical season and in the years that followed.