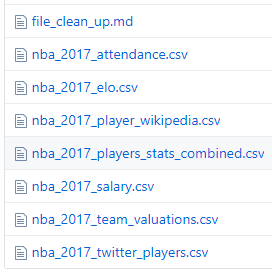
**ETL Project Report**

The social power NBA sports data was extracted, transformed, and loaded in this ETL project. The data have salary, on-court performance, twitter engagement, and Wikipedia traffic. The data was downloaded from Kaggle data sources section at [https://www.kaggle.com/noahgift/social-power-nba#](https://www.kaggle.com/noahgift/social-power-nba).

**Extract**

The data files from Kaggle were in a comma delimited CSV format. The following files were downloaded.



* The attendance data had team names, games played, total attendance, and average attendance per game.
* The ELO data had ratings for each team in both Western and Eastern conferences. ELO is based on the inputs such as score at the end of each game, when and where they are played. One team can gain ELO points by winning against another team that loses points, a zero-sum rating.
* The Wikipedia data had page views with timestamp for each player.
* The combined stats contained player information such as name, age, team, and position played, and statistics such as points, field goals, field goal percentage, free throws etc.
* The salary data had player name, salary, position played, and the team.
* The valuation data had team valuation in millions of dollars for each team.
* The twitter data had favorite and retweet counts for each player.

**Transform**

Below is a visual of the changes that were made to the CSV files that were imported from Kaggle. All final data frames that were loaded into MySQL have been distinguished by their data frame names.

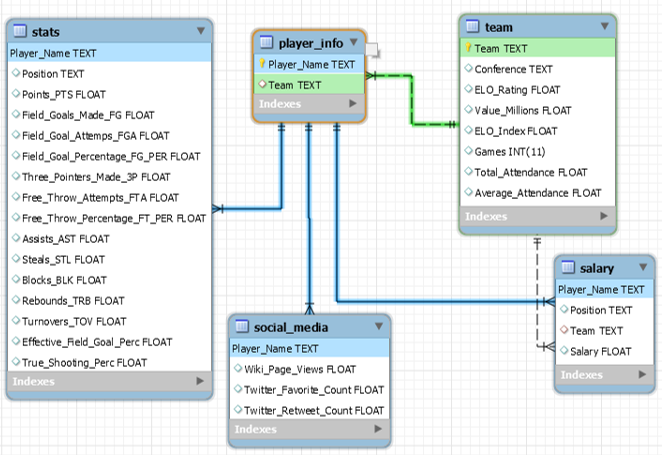
|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  | Related image | **nba\_2017\_social\_influence\_df** |
|  |  |  |
| For players who were listed more than once, we took the average of the salaries listed. |  | **nba\_2017\_salary\_df** |
|  |  |  |
|  |  | **nba\_2017\_player\_df** |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  | Related image |  |
|  |  |  |
|  |  |  |

|  |  |  |
| --- | --- | --- |
|  | Related image |  |
| Calculate ELO\_Index and reorganize columns |  | **nba\_2017\_team\_df** |

|  |
| --- |
|  |
|  |
|  |
| Calculate Effective\_Field\_Goal\_Perc and True\_Shooting\_Perc |
|  |
| **nba\_2017\_stats\_df** |
|  |

**Load**

An entity relationship model and tables were created in a MySQL database, and the data were loaded into the tables from Pandas Dataframes and SQL.



**Queries**

A few SQL queries were created to retrieve data from the social power NBA database that are listed below.

1. **Query Sorted by Salary in Descending Order**

select

i.Player\_Name

,s.Twitter\_Favorite\_Count

,s.Twitter\_Retweet\_Count

,s.Wiki\_Page\_Views

,m.Salary

FROM nba\_social\_power\_db.player\_info i

left outer join nba\_social\_power\_db.social\_media s on i.Player\_Name = s.Player\_Name

left outer join nba\_social\_power\_db.salary m on i.Player\_Name = m.Player\_Name order by m.salary desc;

1. **Query Sorted by Wiki Page Views in Descending Order**

select

i.Player\_Name

,s.Twitter\_Favorite\_Count

,s.Twitter\_Retweet\_Count

,s.Wiki\_Page\_Views

,m.Salary

FROM nba\_social\_power\_db.player\_info i

left outer join nba\_social\_power\_db.social\_media s on i.Player\_Name = s.Player\_Name

left outer join nba\_social\_power\_db.salary m on i.Player\_Name = m.Player\_Name order by s.Wiki\_Page\_Views desc;

1. **Query Sorted by Twitter Favorite Count in Descending Order**

select

i.Player\_Name

,s.Twitter\_Favorite\_Count

,s.Twitter\_Retweet\_Count

,s.Wiki\_Page\_Views

,m.Salary

FROM nba\_social\_power\_db.player\_info i

left outer join nba\_social\_power\_db.social\_media s on i.Player\_Name = s.Player\_Name

left outer join nba\_social\_power\_db.salary m on i.Player\_Name = m.Player\_Name order by s.Twitter\_Favorite\_Count desc;

**Conclusions**

Calculated and added ELO index into teams, Effective Field Goal Percentage and True Shooting Percentage metrics into Player stats. These additional calculations worked to create better performance numbers to compare with different types of data like - social media activity and salary. For a continuous analysis of social power data, we would want to find dynamic data sources, rather than csv files. The next step would be to create scatter plots to see possible relationships.