

THE DATA NEWS BEARS

By: Alyssa Pacleb (ajn873) and Zak Kuzbari (mzk97)

Date: 12-13-2019



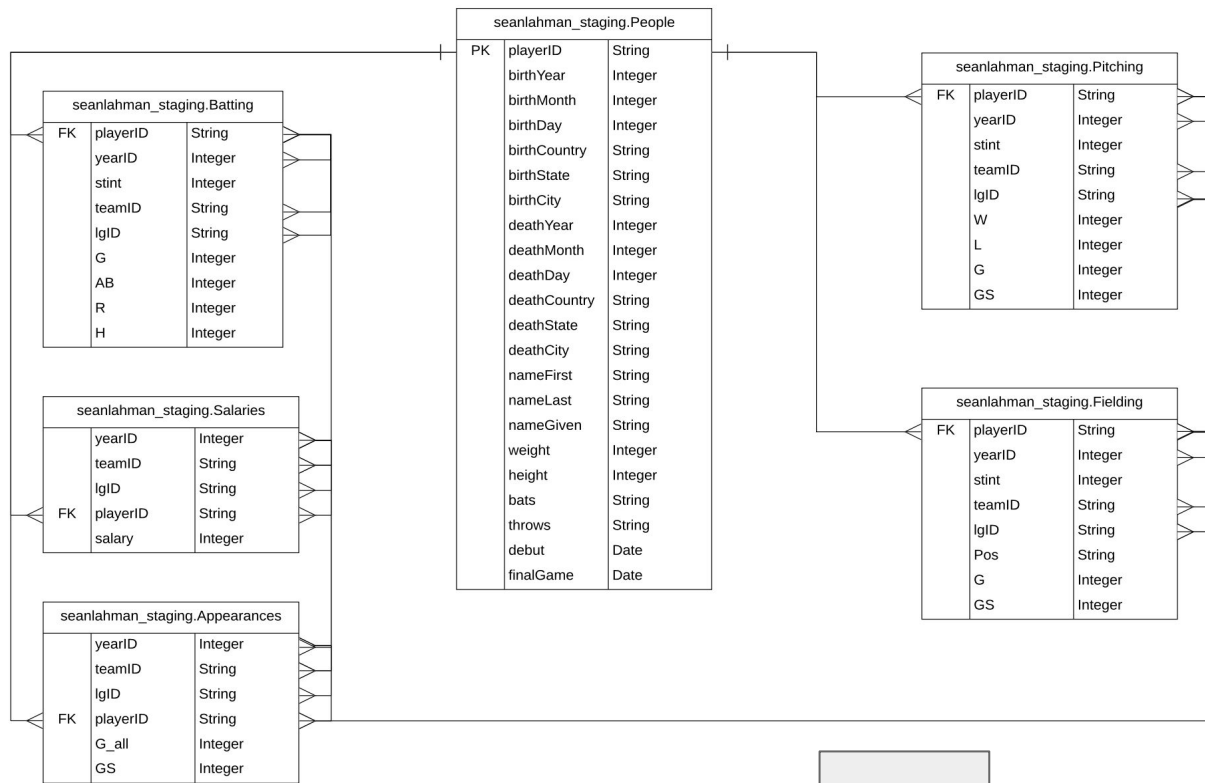
WHY WE CHOSE TOPIC

- “What we assume that the teams should know, but never seem to get, is that you’re paying for the future, not the past.” -Abraham Wyner, Chair of Statistics at Wharton
- More statistics in baseball (MLB) than any other sport
- No salary cap for players in the MLB
- Great Texan franchises (Houston Astros and Texas Rangers)

Datasets:

Major League Baseball vs. Nippon Professional Baseball

INITIAL DATASET



Number of Rows in Datasets

```
select count (*) from
seanlahman_staging.People
--19617
```

```
select count (*) from
seanlahman_staging.Pitching
--46699
```

```
select count (*) from
seanlahman_staging.Appearances
--105789
```

```
select count (*) from
seanlahman_staging.Fielding
--140921
```

```
select count (*) from
seanlahman_staging.Salaries
--26428
```

```
select count (*) from
seanlahman_staging.Batting
--105861
```

INITIAL TRANSFORMS

Player stats all

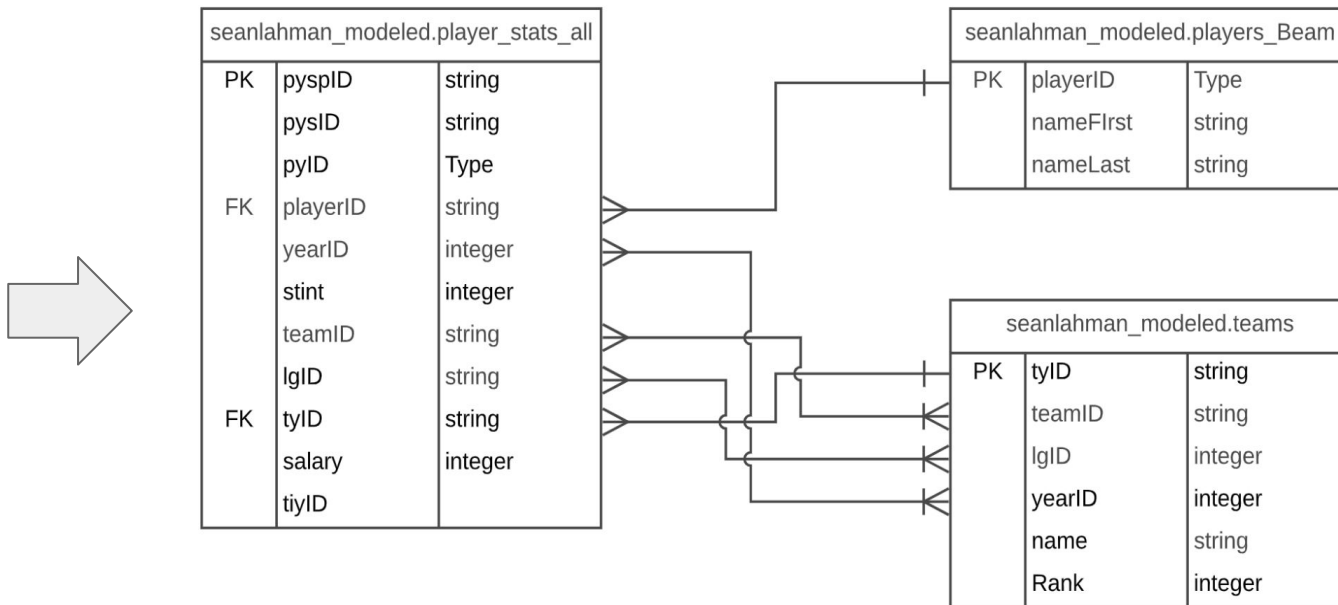
JOIN on playerID for
Batting, Salaries,
Pitching, Fielding.

Players Beam

Transform on *people*
table. Standardized
countries and date of
birth column.

Teams

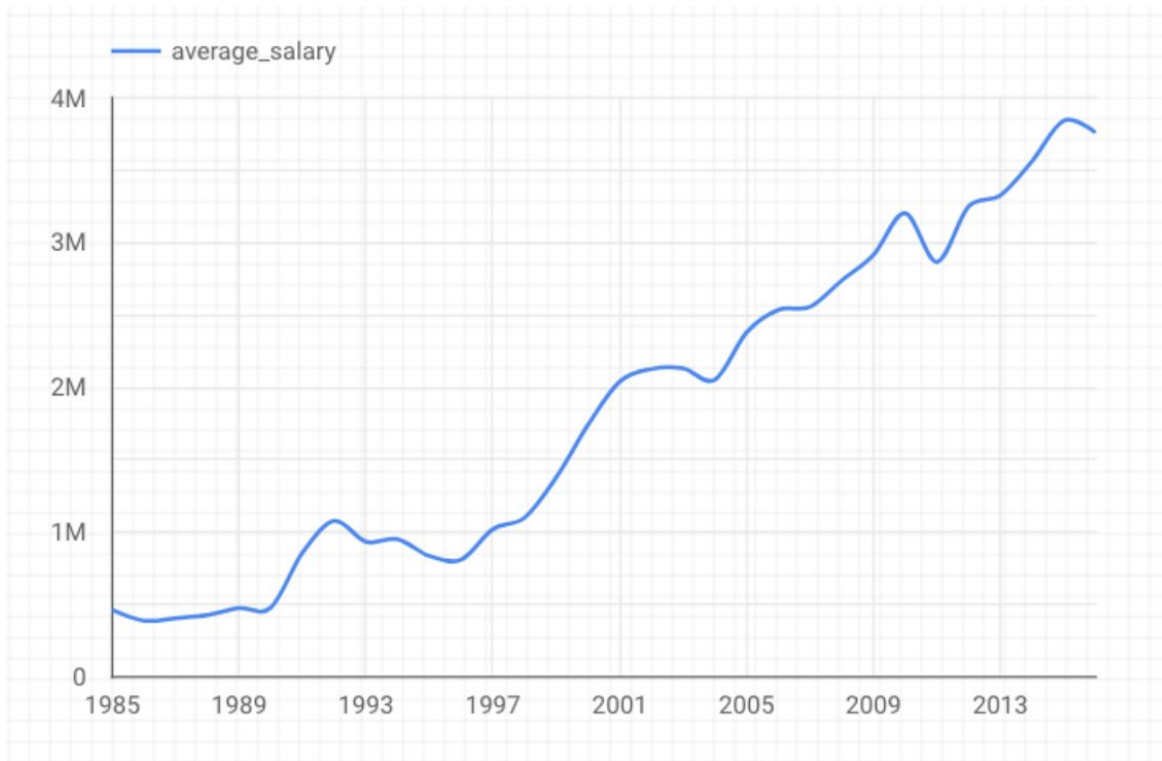
Included later from
original source.



https://github.com/cs327e-fall2019/The-Data-News-Bears/blob/master/players1_cluster.py

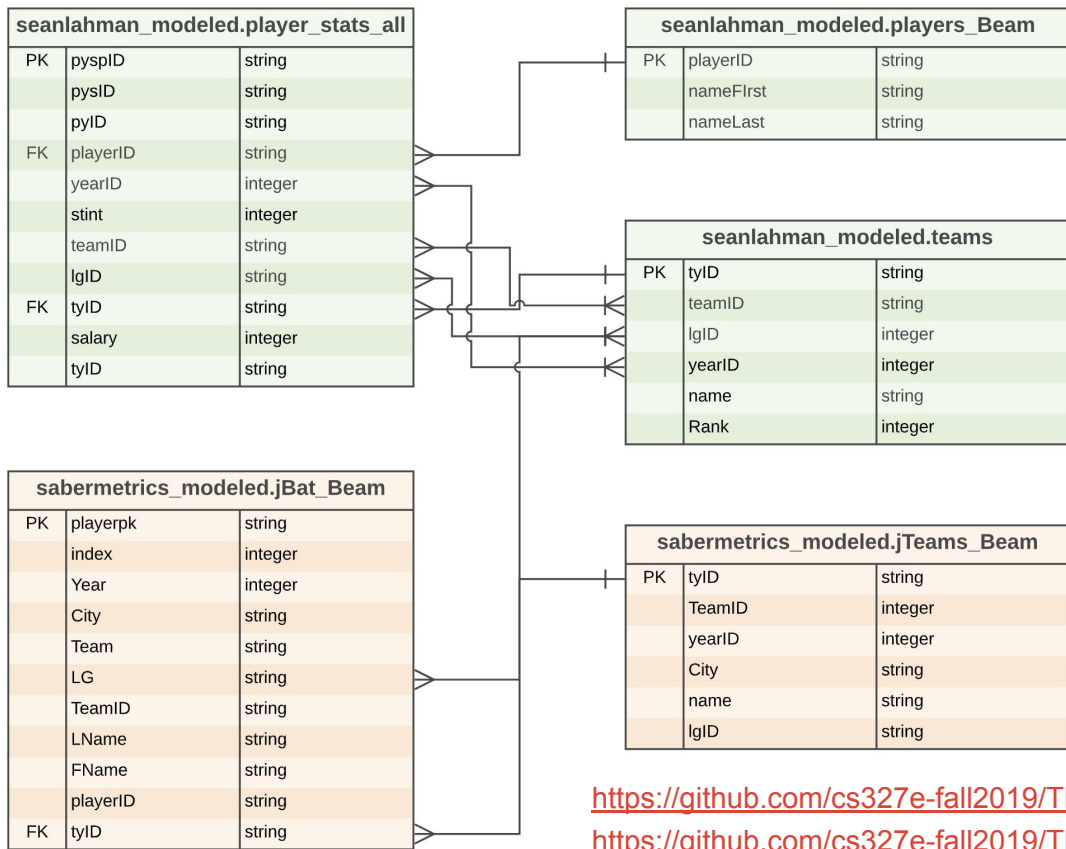
https://github.com/cs327e-fall2019/The-Data-News-Bears/blob/master/players2_cluster.py

INTERESTING QUERIES VISUALISED



Thanks to the MLB's use of the luxury tax instead of a salary cap, players are paid **a lot**

MODELED DATASETS: BEAM + DAG TRANSFORMS



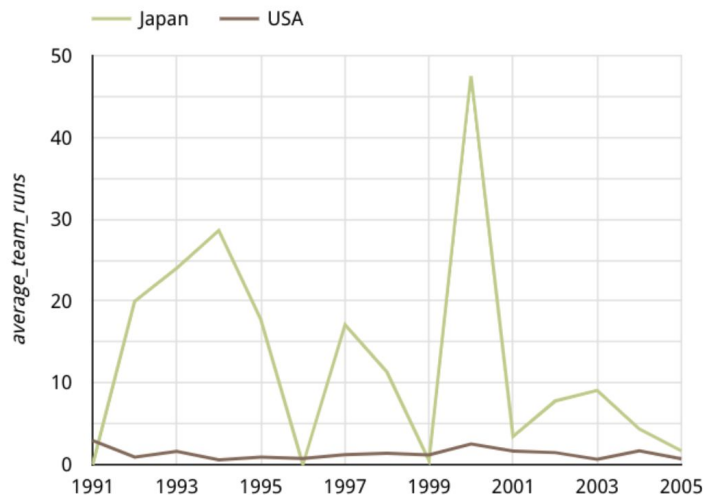
- Created the jTeams_Beam table to average team statistics
- Standardized team and city names and created primary keys for each table
- DAG made to produce same results as beam transforms and sql queries



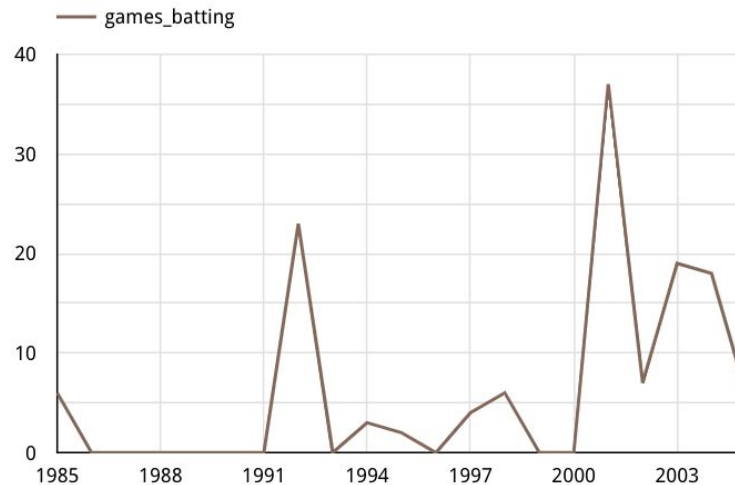
https://github.com/cs327e-fall2019/The-Data-News-Bears/blob/master/transform_jBat_single.ipynb

https://github.com/cs327e-fall2019/The-Data-News-Bears/blob/master/sabermetrics_workflow.ipynb

VISUALIZING THE DIFFERENCES

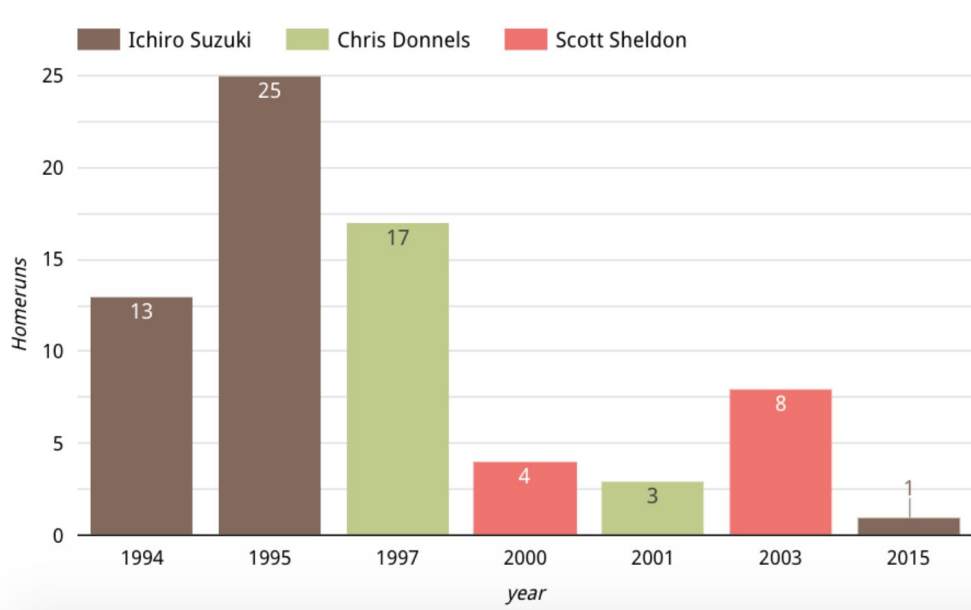


- The Japanese teams consistently scored more points than the MLB teams.



- In recent years, Japanese players have American counterparts that bat equal number of times

VISUALIZING THE DIFFERENCES: DEMO



- Ichiro Suzuki and Chris Donnels went from the NPB to the MLB did very poorly after the transition
- Scott Sheldon, an American player who went from the MLB to the NPB, did a lot better

INSIGHTS ABOUT BASEBALL (MLB IS “BETTER”)

- The way that each country's league is trained is evident in their strengths.
- According to our visualizations, why does the Japanese league seem “easier” to play in?

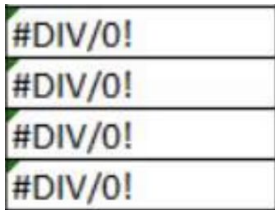
Differences from Major League Baseball



The rules are essentially those of [Major League Baseball](#) (MLB), but technical elements are slightly different: The Nippon league uses a smaller baseball, strike zone, and playing field. Five Nippon league teams have fields whose small dimensions would violate the American [Official Baseball Rules](#).^[6]

ISSUES AND NEXT STEPS FOR IMPROVEMENT

- Creator of Japanese League data most likely used excel to calculate some column values, so the validity of the data is questionable



#DIV/0!
#DIV/0!
#DIV/0!
#DIV/0!

- Learn how to scrape data from non-programmer friendly sites: <https://www.baseball-reference.com/register/npb-stats.shtml>
- Could've paid for better baseball data - looked at height, weight, and salaries for Japanese players

Questions?