The Data News Bears

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

Date: 12-06-2019

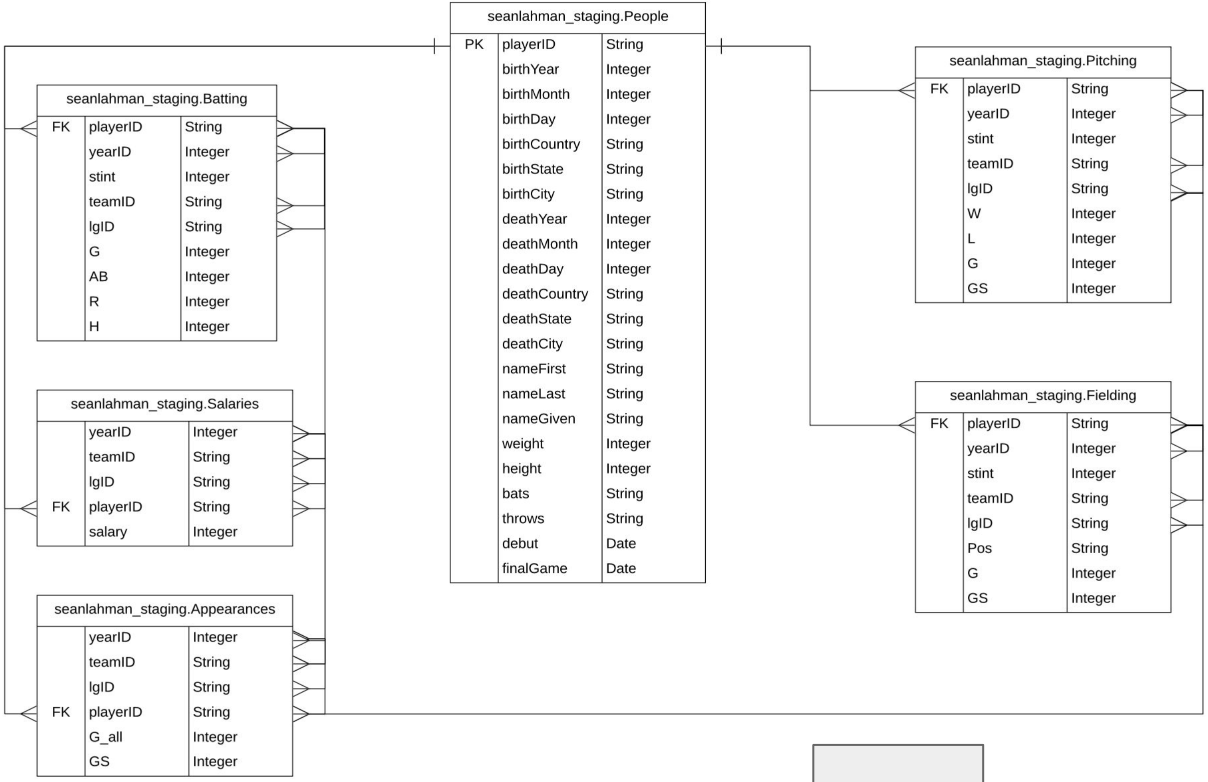
Why We chose topic

- “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 datasets



seanlahman_staging.People

PK	playerID	String
	birthYear	Integer
	birthMonth	Integer
	birthDay	Integer
	birthCountry	String
	birthState	String
	birthCity	String
	deathYear	Integer
	deathMonth	Integer
	deathDay	Integer
	deathCountry	String
	deathState	String
	deathCity	String
	nameFirst	String
	nameLast	String
	nameGiven	String
	weight	Integer
	height	Integer
	bats	String
	throws	String
	debut	Date
	finalGame	Date

seanlahman_staging.Batting

FK	playerID	String
	yearID	Integer
	stint	Integer
	teamID	String
	lgID	String
	G	Integer
	AB	Integer
	R	Integer
	H	Integer

seanlahman_staging.Salaries

	yearID	Integer
	teamID	String
	lgID	String
FK	playerID	String
	salary	Integer

seanlahman_staging.Appearances

	yearID	Integer
	teamID	String
	lgID	String
FK	playerID	String
	G_all	Integer
	GS	Integer

seanlahman_staging.Pitching

FK	playerID	String
	yearID	Integer
	stint	Integer
	teamID	String
	lgID	String
	W	Integer
	L	Integer
	G	Integer
	GS	Integer

seanlahman_staging.Fielding

FK	playerID	String
	yearID	Integer
	stint	Integer
	teamID	String
	lgID	String
	Pos	String
	G	Integer
	GS	Integer

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

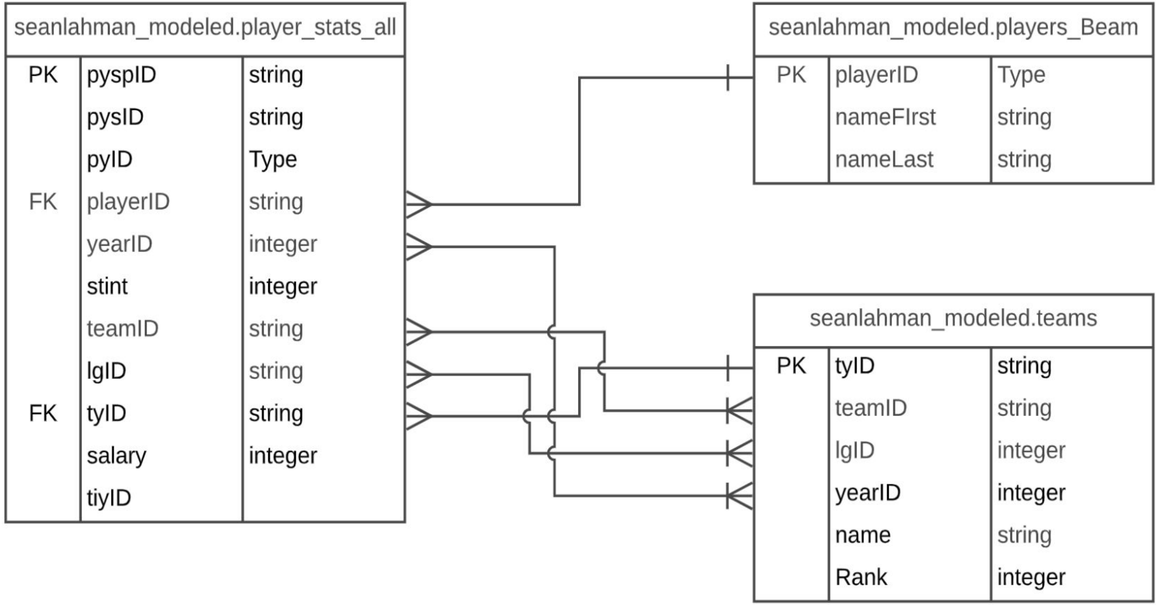
JOIN on playerID for Batting, Salaries, Pitching, Fielding.

Players Beam

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

Teams

Included later from original source.



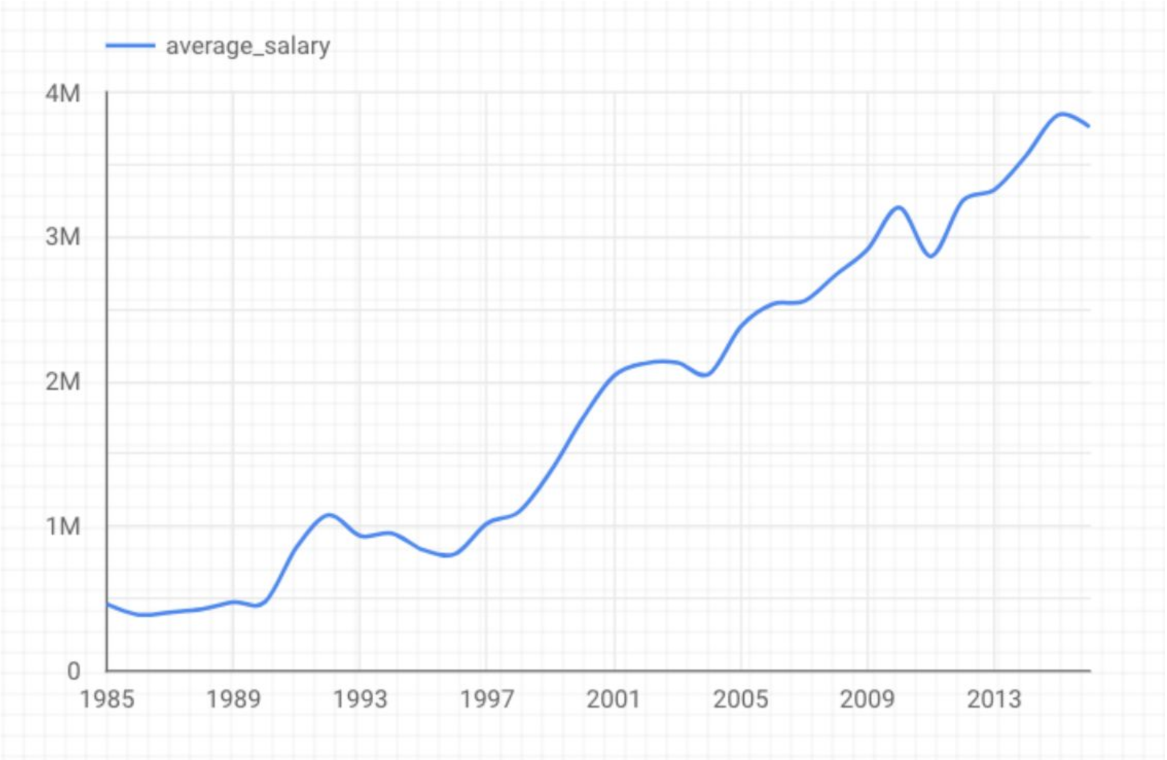
seanlahman_modeled.player_stats_all		
PK	pyspID	string
	pysID	string
	pyID	Type
FK	playerID	string
	yearID	integer
	stint	integer
	teamID	string
	lgID	string
FK	tyID	string
	salary	integer
	tiyID	

seanlahman_modeled.players_Beam		
PK	playerID	Type
	nameFIrst	string
	nameLast	string

seanlahman_modeled.teams		
PK	tyID	string
	teamID	string
	lgID	integer
	yearID	integer
	name	string
	Rank	integer

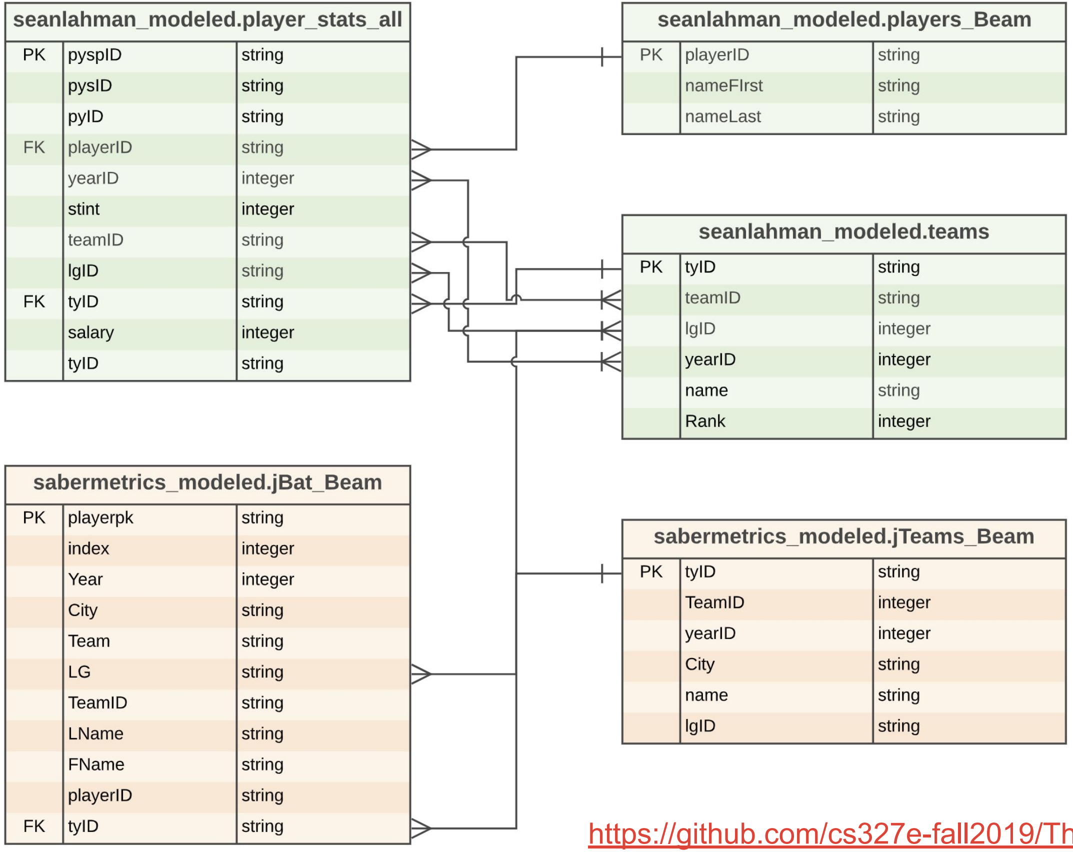
https://github.com/cs327e-fall2019/The-Data-News-Bears/blob/master/players1_cluster.py
https://github.com/cs327e-fall2019/The-Data-News-Bears/blob/master/players2_cluster.py

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



seanlahman_modeled.player_stats_all

PK	pyspID	string
	pysID	string
	pyID	string
FK	playerID	string
	yearID	integer
	stint	integer
	teamID	string
	lgID	string
FK	tyID	string
	salary	integer
	tyID	string

seanlahman_modeled.players_Beam

PK	playerID	string
	nameFIrst	string
	nameLast	string

seanlahman_modeled.teams

PK	tyID	string
	teamID	string
	lgID	integer
	yearID	integer
	name	string
	Rank	integer

sabermetrics_modeled.jBat_Beam

PK	playerpk	string
	index	integer
	Year	integer
	City	string
	Team	string
	LG	string
	TeamID	string
	LName	string
	FName	string
	playerID	string
FK	tyID	string

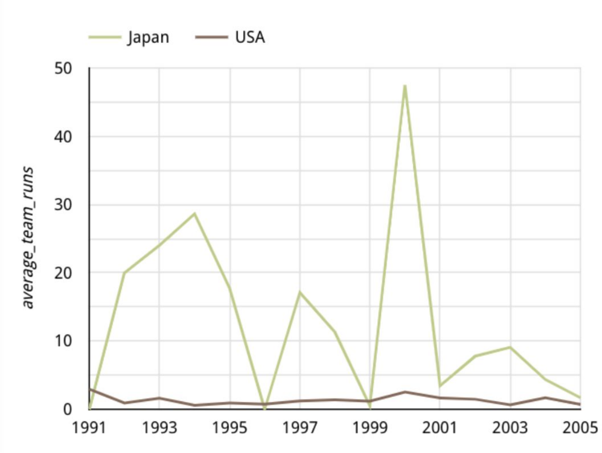
sabermetrics_modeled.jTeams_Beam

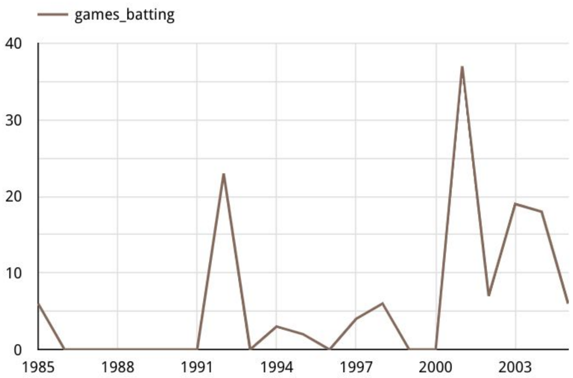
PK	tyID	string
	TeamID	integer
	yearID	integer
	City	string
	name	string
	lgID	string

- 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

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

Visualizing the differences

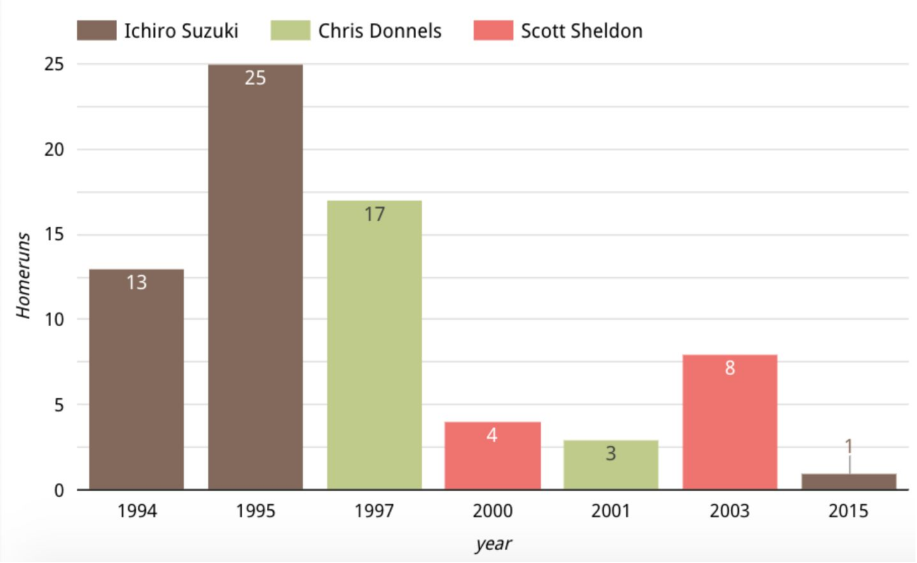




- The Japanese teams consistently scored more points than the MLB teams.

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

Visualizing the differences: Demo



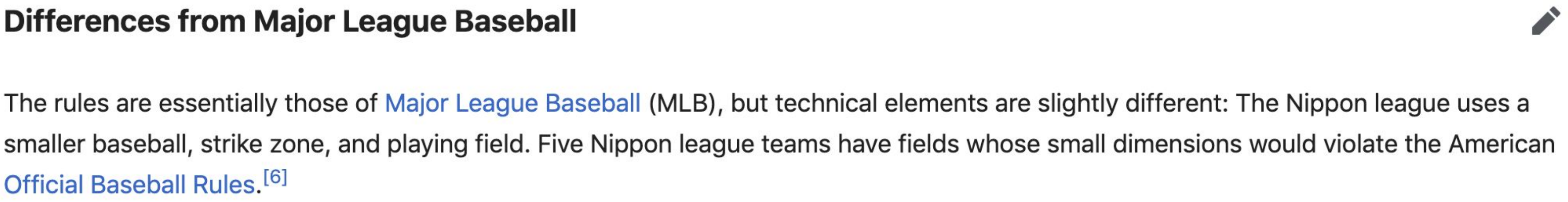
- Ichiro Suzuki and Chris Donnels went from the NPB to the MLB did very poorly after the transition
- Scott Sheldon, an American player who went from the MLB to the NPB, did a lot better

https://console.cloud.google.com/bigquery?sq=761926234191:8ac2f4c3d67042a59b79b5fe1238e6e1

COOL 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?

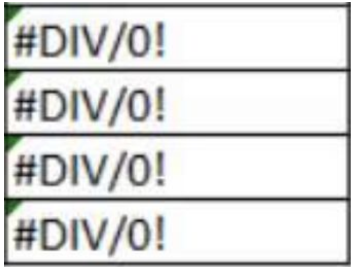


Differences from Major League Baseball

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

Questions?