PROJECT 1
DATA VISUALIZATION

Submitted By

Pranesh Narayanan
Yuvanesh Vedaraju

Contents

Data Source

We obtained our data used, from Lahmans Database
Reference Link: http://www.seanlahman.com/baseball-archive/statistics/

Files Used:

File 1: Master Data

Player Information from player id, player name, dates of birth and other demographical info

Meta Data

Player Id: It is a unique identification number that each player possesses.
Age: Calculated field that displays the age of a given player by year

```
masterTemp <- Master[, c('playerID', 'birthYear', 'birthMonth',
                         'nameLast', 'nameFirst', 'bats')]

Salaries <- merge(Salaries, masterTemp, all.x = TRUE)

Salaries$age <- with(Salaries, yearID - birthYear -
                       ifelse(birthMonth < 10,0,1))
```

Year Id: Year in which statistics linked to occurred
Team Id: It is a unique identification number that every team possesses
Lg Id: a factor with levels AL NL
Birth Year: Year in which given player was born
Birth Month: Month in which given player was born
Birth Date: Date in which given player was born
Birth Country: Country in which given player was born
Birth State: State in which given player was born
First Name: A player's first name
Last Name: A player's last name
Weight: A player's weight in pounds
Height: A player's height in inches
Bats: whether the player is left handed or right handed
Throws: whether the player throws with his left hand or right hand

File 2: Salaries

This file consists of the salary information of players
Player Id: It is a unique identification number that each player possesses.
Team Id: It is a unique identification number that every team possesses
Lg Id: a factor with levels AL NL
Year Id: Year in which statistics linked to occurred
Salary: Salary of players

File 3: Batting Table

The batting table which contained batting statistics
Player Id: It is a unique identification number that each player possesses.
Team Id: It is a unique identification number that every team possesses

Lg Id: a factor with levels AL NL
R (Runs): No. runs of scored by the player
H: Hits
2b: Doubles
3b: Triples
HR: Home runs
Ab: At bats
BB: Walks

File 4: Pitching Table

Pitching table which contained pitching stats
Player Id: It is a unique identification number that each player possesses.
Team Id: It is a unique identification number that every team possesses
Lg Id: a factor with levels AL NL
Age: Calculated field that displays the age of a given player by year
Year Id: Year in which statistics linked to occurred
ERA: Earned Run Average
POS: Position Played

File 5: Fielding

Player Id: It is a unique identification number that each player possesses.
Team Id: It is a unique identification number that every team possesses
POS: Position Played

Combining Dataset:

Steps followed

We used R to restructure and manipulate our data, merge our data sets(playerID was used as Primary Key), exempt players with no salary information(year <= 1985), create certain variables such as age:

```
masterTemp <- Master[, c('playerID', 'birthYear', 'birthMonth', 'nameLast', 'nameFirst', 'bats')]

Salaries <- merge(Salaries, masterTemp, all.x = TRUE)

Salaries$age <- with(Salaries, yearID - birthYear -

                     ifelse(birthMonth < 10,0,1))
```

and reorganize our data sets by pitchers and fielders/batters.

The merged file was used as our primary file and the duplicate columns are removed

Joining Files:

We used Analysis_R.xlsx to create new files named as SalaryPitchingNewest and SalaryBattingNewest which we used for analysis in Tableau.
These files contain a time series data from 1985 to 2014 for each player with over 1100 records.

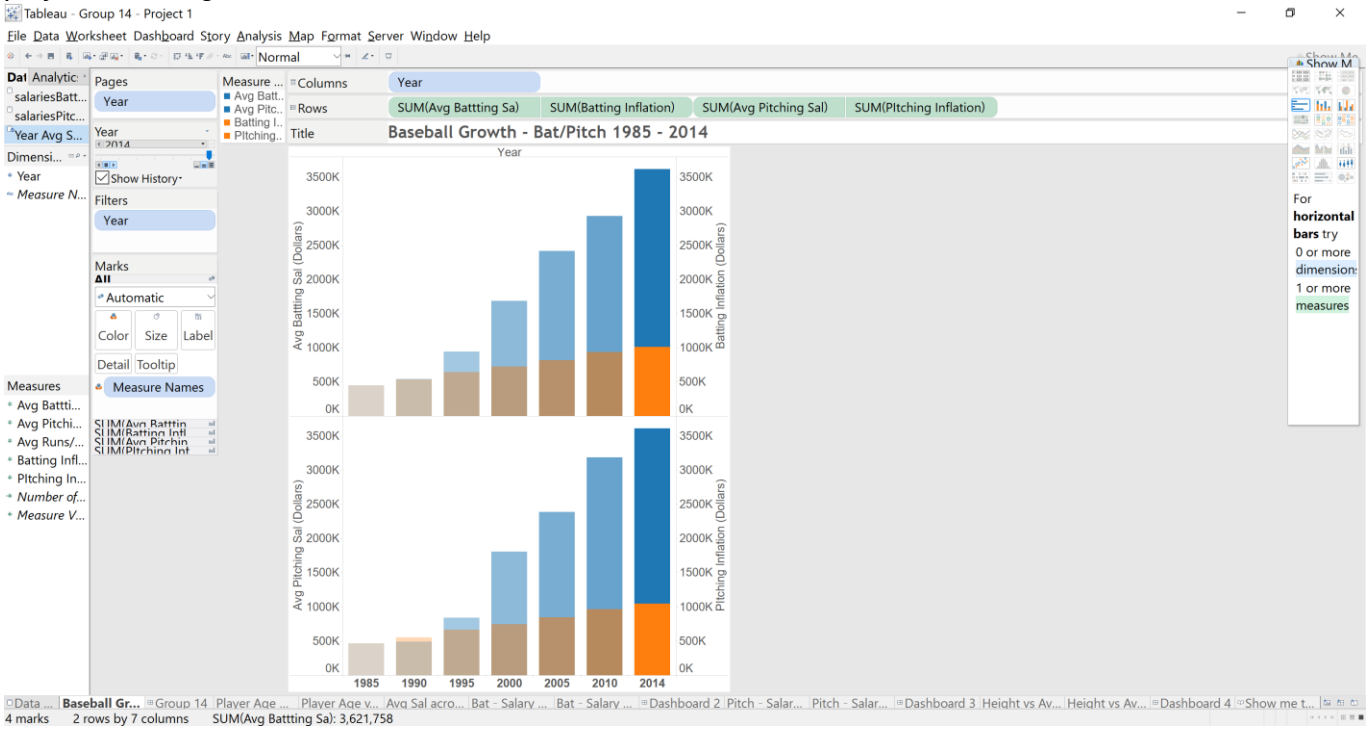
After transforming our data, the next step was exploratory analysis in which we sought to discover insights into what factors lead to higher salaries.
As we got our joined files we have uploaded it on Tableau to do the analysis.

Calculated Fields:

1) Slugging: A baseball statistic that is calculated by determining the total number of bases a batter has reached divided by their total number of At-Bats.
2) No. Of players: Number of records needed to be slightly modified for our requirements. We needed to know unique players across various attributes so we used CountD function against player id.

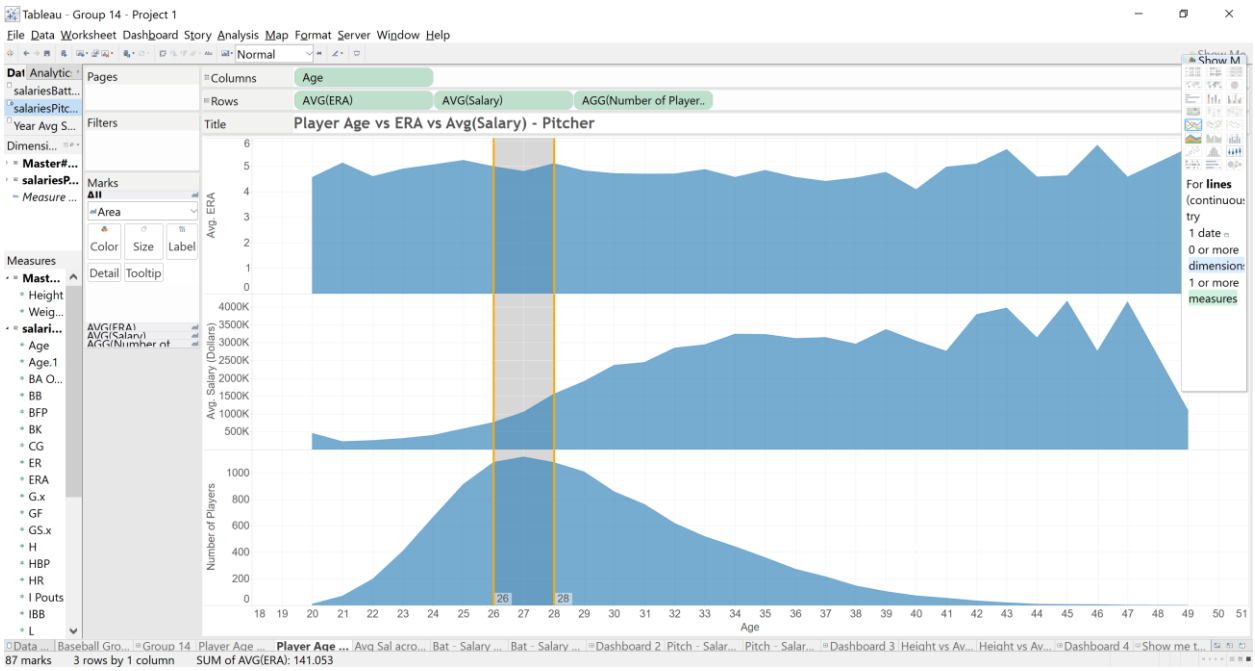
Insights:

1. We discovered that the average baseball player's salary grew faster than the rate of inflation, unlike many US professions, suggesting that there was a very high demand for good players, competition for player acquisition was heavy, or that the industry as a whole was becoming more profitable. We believe all of these factors had a say in this rise above inflation, but given the media's rise to popularity in relation with live professional sporting events, so ad revenues likely experienced drastic spikes. This increase in media also likely contributed to the 'branding' of players, working in the favor of their salaries as well.



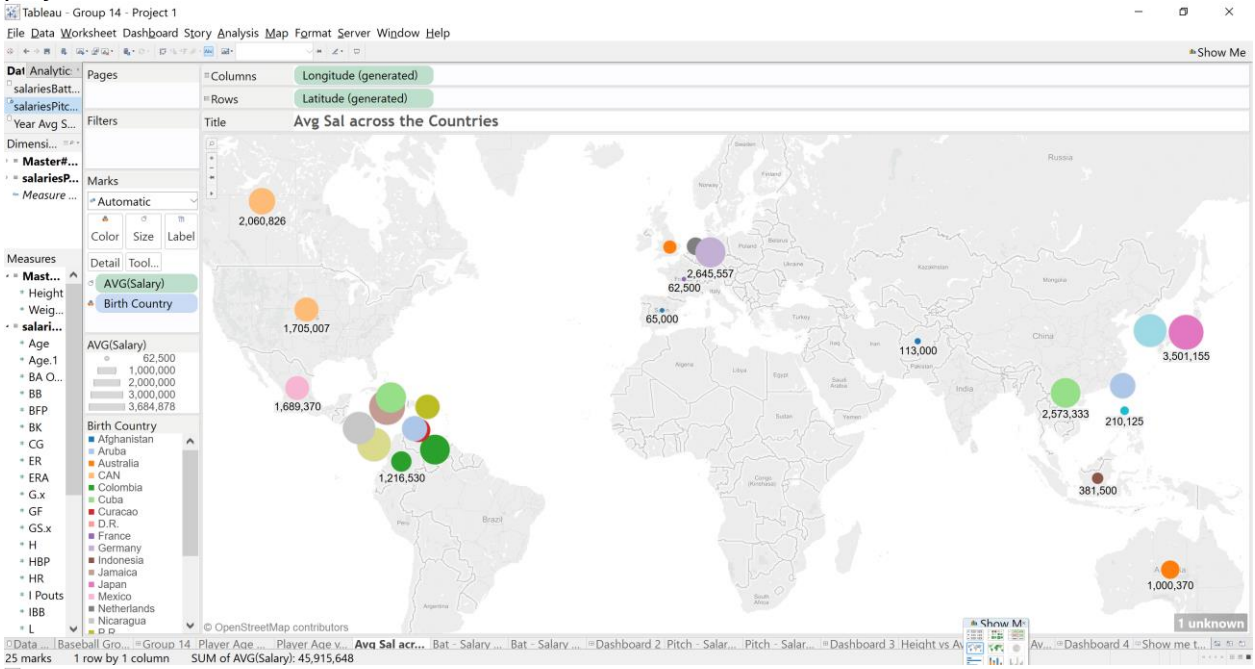
Also noteworthy, was the salaries shared by pitchers, and batters, collectively as groups did not deviate from one another. This was unexpected to us, as we were looking to find a pattern between pitcher/batter performance and their respective compensation. This simply was not the case. However, we would not rule out the possibility that these patterns exist further back in history, but we were unable to further pursue this given the limitation of our data.

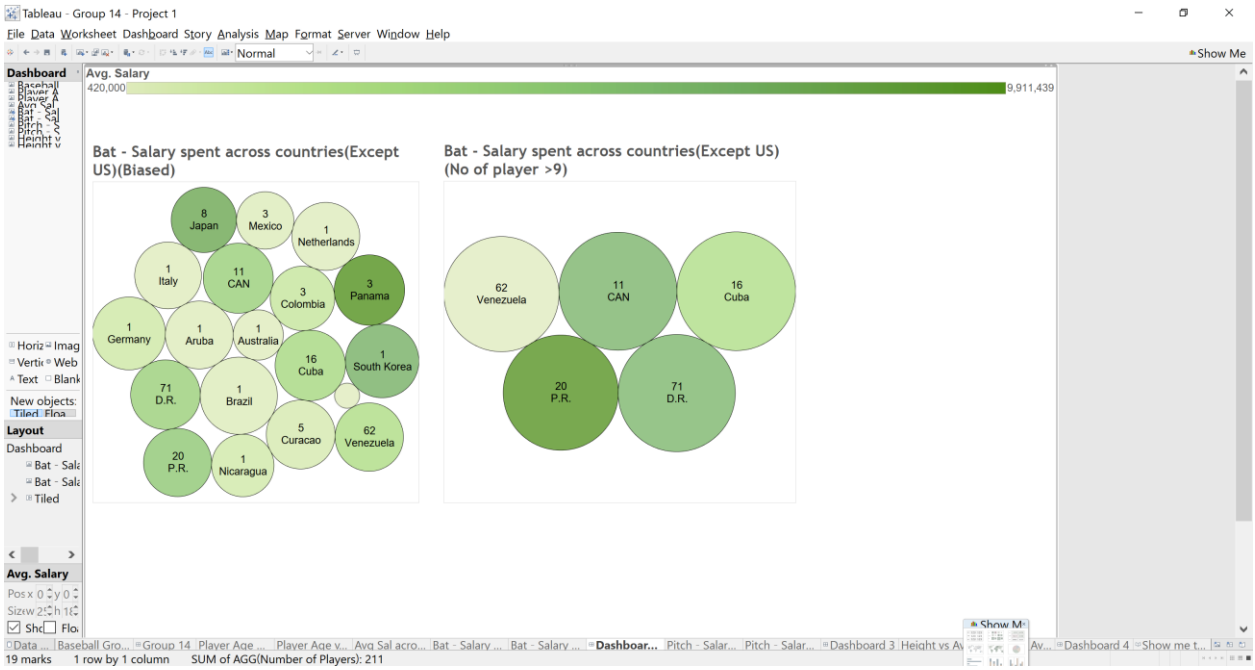
2. When looking at age and salary, we noticed that salaries tend to begin rising heavily after the age of 27, and would typically stay rather steady until retirement.



Ironically, we observe that the performance of the players (measured against Slug percentage) rather stays steady across age, also considering the low-scale effect.
From this we can concur that, the teams would be better off investing on new players rather than retaining existing old players with exorbitant salary pay without affecting the success-factor
Note :- In the rare case that the player remain in the league an abnormally long time, we saw that salaries were adjusted (likely signaling a new contract negotiation).

3. When looking at nationality and salary, the correlation for the foreign born players varied, but US players did not have the highest average salary, likely because the incentive to signing foreign players often implies more talent, while US players will feature a much broader talent pool, since less barriers exist for them and there is significantly more access to potentially sign American players.





However, it was note-worthy to observe that the salary paid for players from different countries varied a lot.
From the above, we can see that –

Batters – Although Japan and Columbia have the same amount of Slug percentage, Mexico players on an average are paid very less compared to Japan.

Pitchers – In this case too, Japan is highly paid compared to Columbia, even though the ERA is same across 2 the countries

We created an unbiased version, because the number of players from each country is required to determine if the case is an outlier or not. Hence we filter the countries which have contributed at least 9 players to MLB

4. When looking at height and salary, we found a strong correlation in taller players and higher salaries. This speaks to the nature of the game more than anything else, as taller batter typically, and stressing typically, hit with more power. Also, taller pitchers would be thought to also have wingspans that measure accordingly, which impact a player's pitch velocity as well as general control.