Baseball Analysis

Created by: Ali Krakowsky

Summary

How can historical baseball statistics help us with baseball predictions?

Summary cont.

1. Where do professional players come from?
2. Should the National League adopt the DH rule?
3. What caused the 1990's spike in pitching era and batting average?
4. Does professional experience improve fielding percentage?

How to Answer

1. Jupyter Notebook, Pandas, MatplotLib and Numpy
2. CSV files from Kaggle: The History of Baseball
    1. Fielding
    2. Pitching
    3. Batting
    4. Players
3. Online history of baseball

Data Cleanup

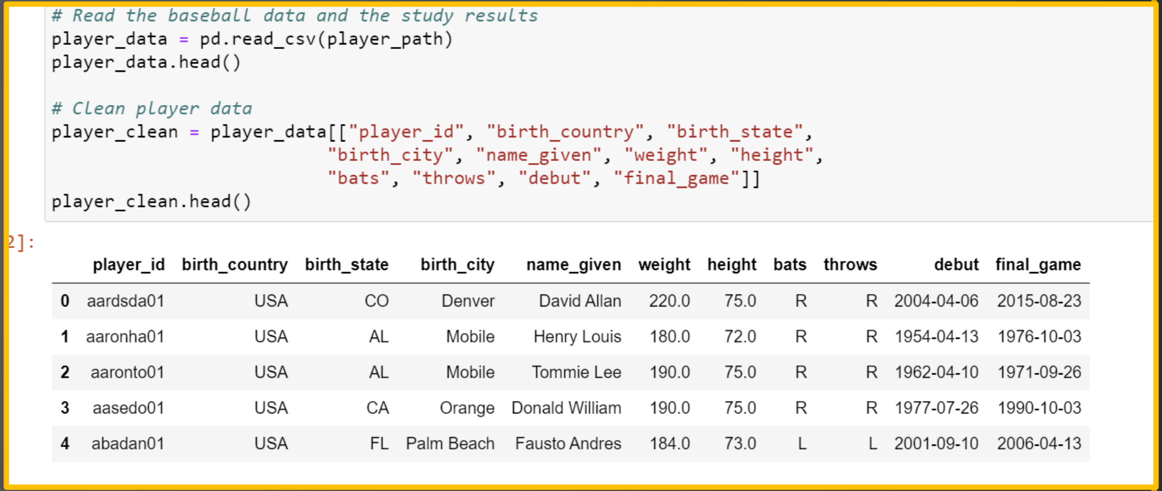
- Source provided 29 csv files with a total of 704 columns
- Cleanup process:
  1. Select csv files
  2. Determine necessary columns
  3. Focus on US born for location
  4. Selected AL and NL leagues
  5. Limited data to more recent years
  6. When combining files selected appropriate field to merge on

Set-backs

1. Deciding which data to use and what questions to ask
2. Individual player data could be separated for the same year if different position and/or team
3. Outside factors could influence stats

Analysis

1. Cleaning up the player data
   - Reduced columns



```python
# Read the baseball data and the study results
player_data = pd.read_csv(player_path)
player_data.head()

# Clean player data
player_clean = player_data[["player_id", "birth_country", "birth_state",
                            "birth_city", "name_given", "weight", "height",
                            "bats", "throws", "debut", "final_game"]]
player_clean.head()
```

2]:

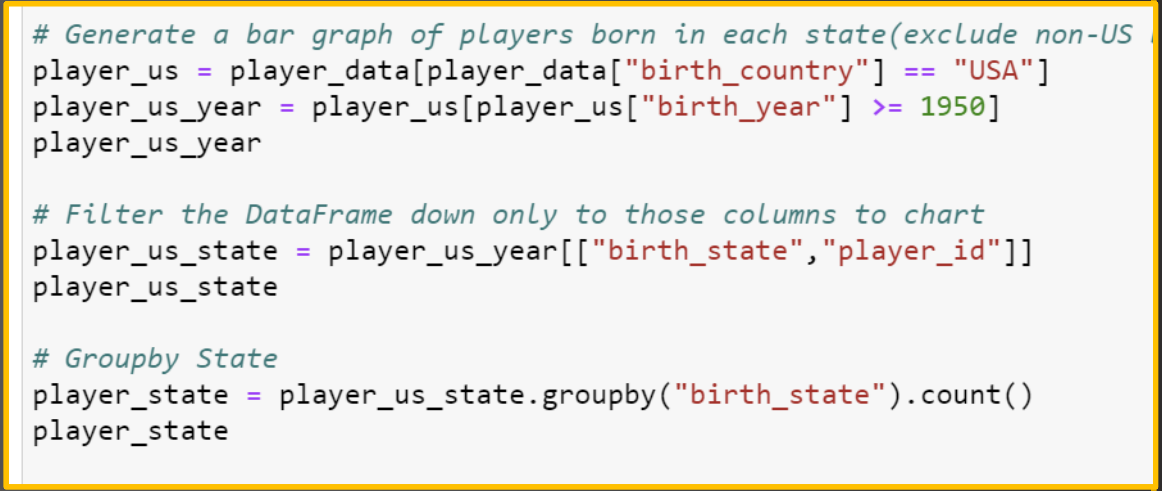
	player_id	birth_country	birth_state	birth_city	name_given	weight	height	bats	throws	debut	final_game
0	aardsda01	USA	CO	Denver	David Allan	220.0	75.0	R	R	2004-04-06	2015-08-23
1	aaronha01	USA	AL	Mobile	Henry Louis	180.0	72.0	R	R	1954-04-13	1976-10-03
2	aaronto01	USA	AL	Mobile	Tommie Lee	190.0	75.0	R	R	1962-04-10	1971-09-26
3	aasedo01	USA	CA	Orange	Donald William	190.0	75.0	R	R	1977-07-26	1990-10-03
4	abadan01	USA	FL	Palm Beach	Fausto Andres	184.0	73.0	L	L	2001-09-10	2006-04-13

Analysis

2. Where are players from?

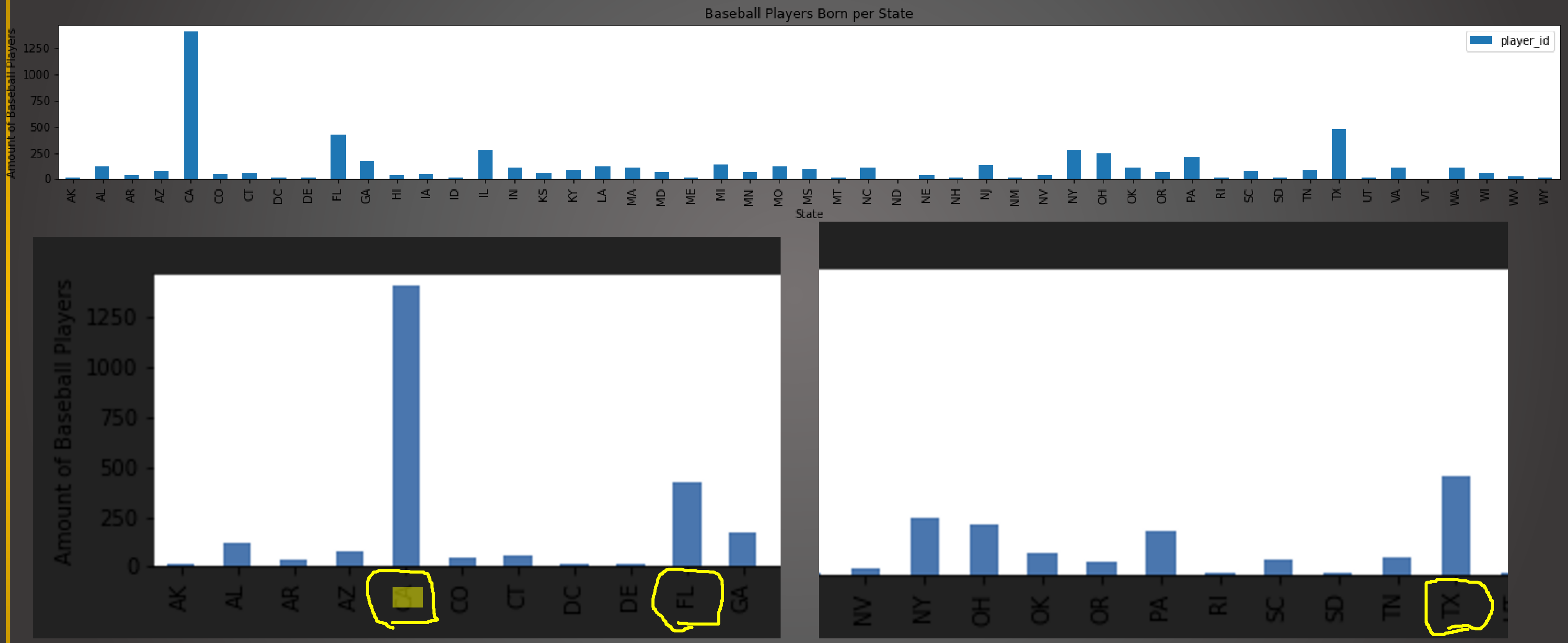
Focused on players born after 1950

US born only



```
# Generate a bar graph of players born in each state(exclude non-US
player_us = player_data[player_data["birth_country"] == "USA"]
player_us_year = player_us[player_us["birth_year"] >= 1950]
player_us_year

# Filter the DataFrame down only to those columns to chart
player_us_state = player_us_year[["birth_state","player_id"]]
player_us_state

# Groupby State
player_state = player_us_state.groupby("birth_state").count()
player_state
```

Figure 1

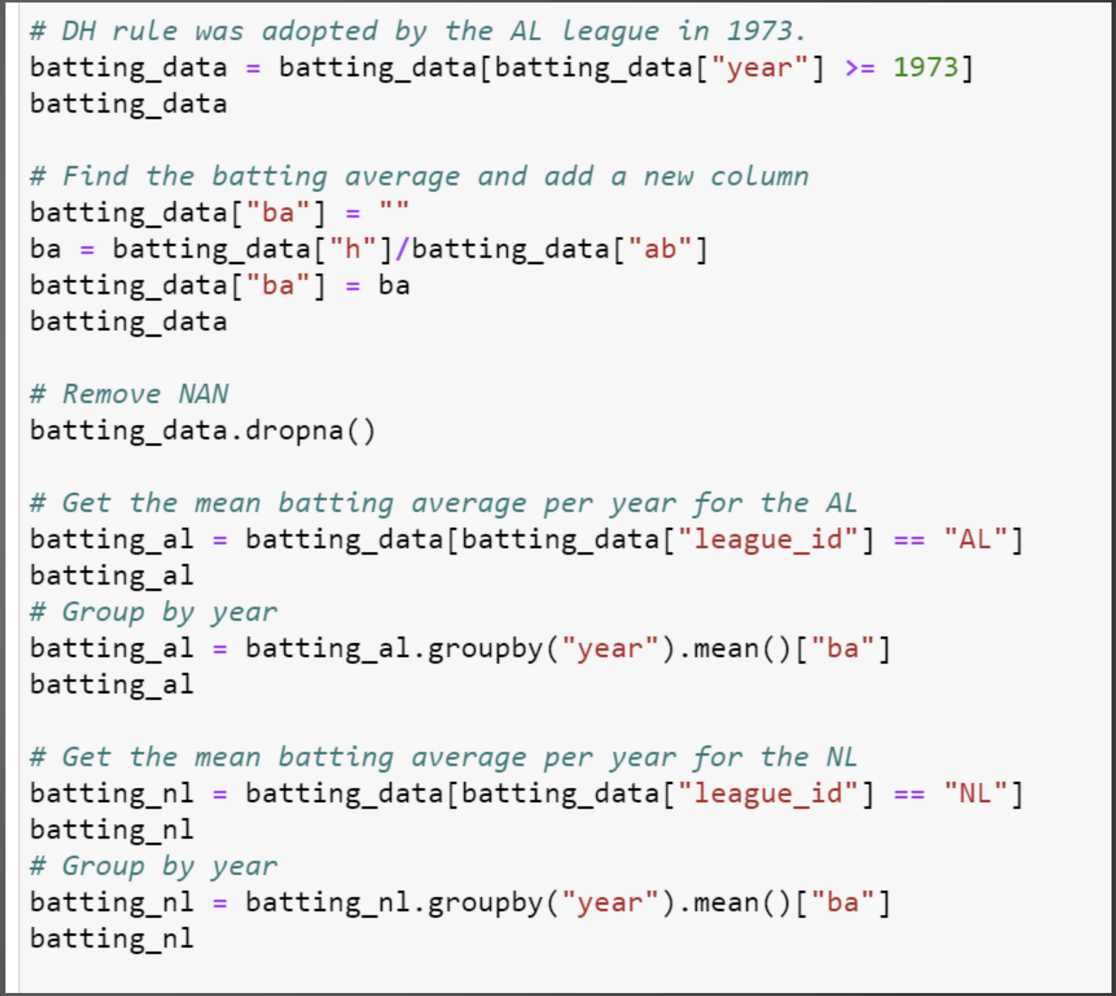

Baseball Players Born per State
Amount of Baseball Players
0
250
500
750
1000
1250
player_id
AK
AL
AR
AZ
CA
CO
CT
DC
DE
FL
GA
HI
IA
ID
IL
IN
KS
KY
LA
MA
MD
ME
MI
MN
MO
MS
MT
NC
ND
NE
NH
NJ
NM
NV
NY
OH
OK
OR
PA
RI
SC
SD
TN
TX
UT
VA
VT
WA
WI
WV
WY
State
Amount of Baseball Players
1250
1000
750
500
250
0
AK
AL
AR
AZ
CO
CT
DC
DE
FL
GA
NV
NY
OH
OK
OR
PA
RI
SC
SD
TN
TX

Analysis

3. Should the NL adopt the DH rule?

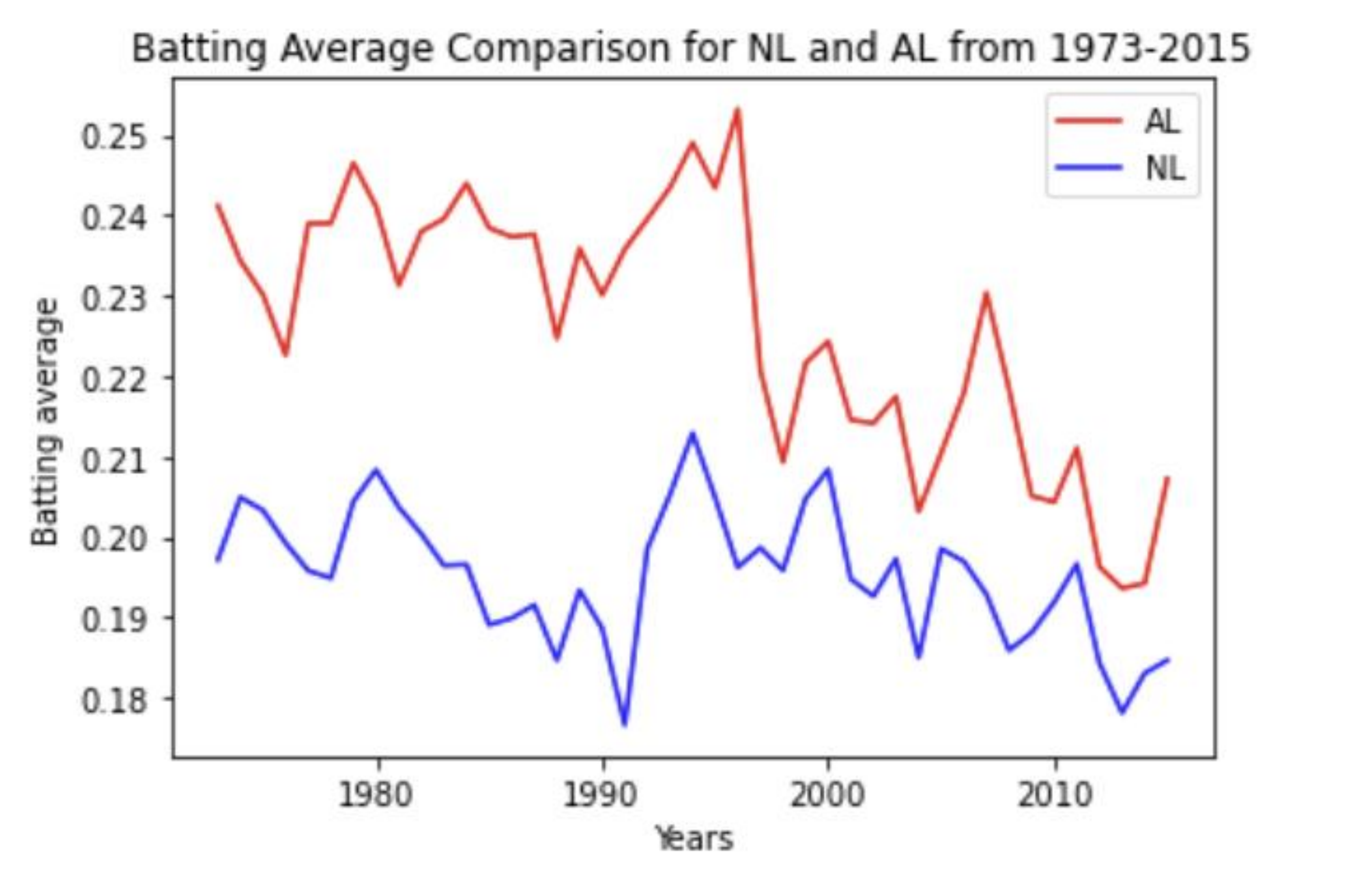
DH rule adopted by AL in 1973

- Data from 1973-2015
- Grouped by each league and found the average batting average per year
- Batting average = hits/at bats



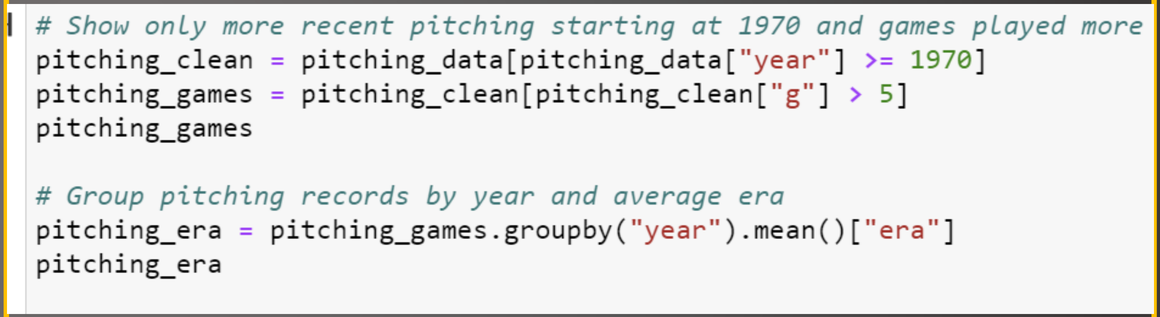
```
# DH rule was adopted by the AL league in 1973.
batting_data = batting_data[batting_data["year"] >= 1973]
batting_data

# Find the batting average and add a new column
batting_data["ba"] = ""
ba = batting_data["h"]/batting_data["ab"]
batting_data["ba"] = ba
batting_data

# Remove NAN
batting_data.dropna()

# Get the mean batting average per year for the AL
batting_al = batting_data[batting_data["league_id"] == "AL"]
batting_al
# Group by year
batting_al = batting_al.groupby("year").mean()["ba"]
batting_al

# Get the mean batting average per year for the NL
batting_nl = batting_data[batting_data["league_id"] == "NL"]
batting_nl
# Group by year
batting_nl = batting_nl.groupby("year").mean()["ba"]
batting_nl
```

Figure 2


Batting Average Comparison for NL and AL from 1973-2015
AL
NL
Batting average
0.25
0.24
0.23
0.22
0.21
0.20
0.19
0.18
1980
1990
2000
2010
Years

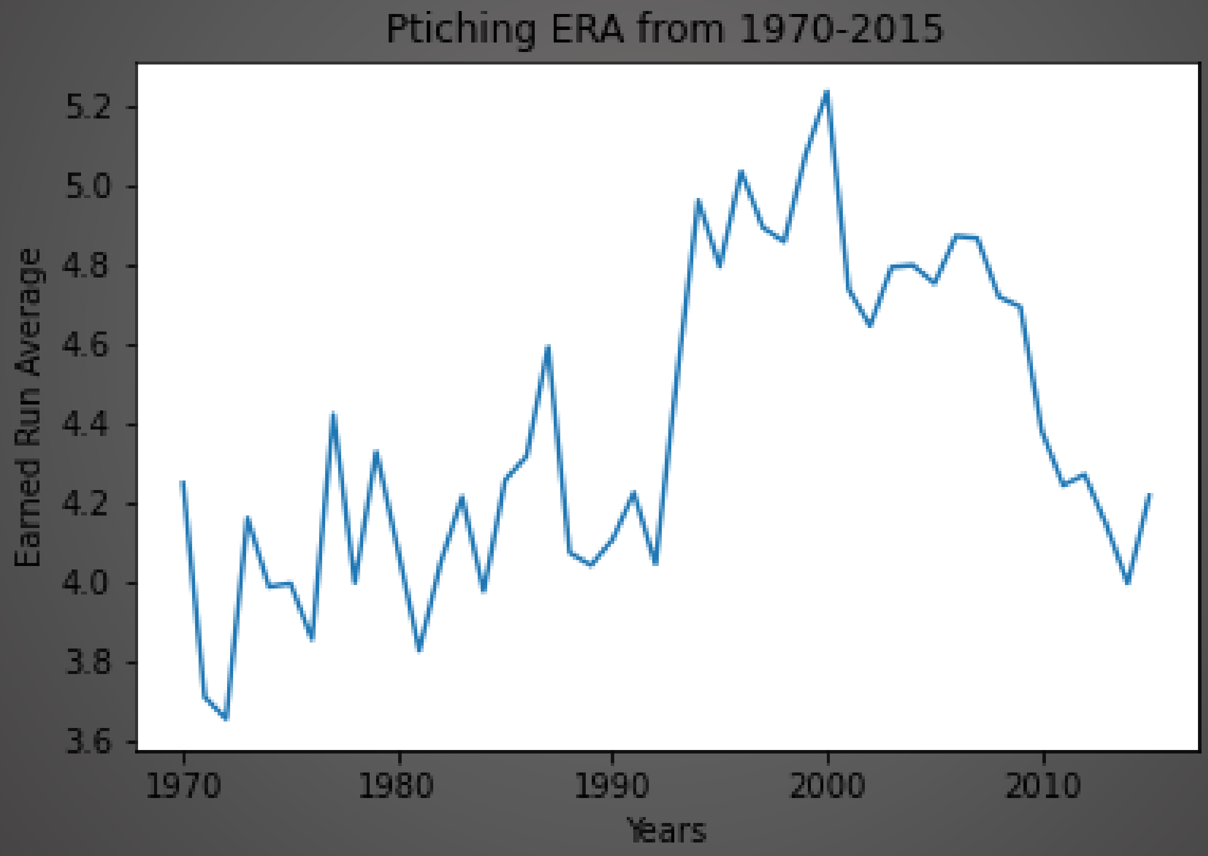
Analysis

4. Has pitching era improved?
   - Years 1970-2015
   - More than 5 games pitched/year
   - Average era per year



```
# Show only more recent pitching starting at 1970 and games played more
pitching_clean = pitching_data[pitching_data["year"] >= 1970]
pitching_games = pitching_clean[pitching_clean["g"] > 5]
pitching_games

# Group pitching records by year and average era
pitching_era = pitching_games.groupby("year").mean()["era"]
pitching_era
```

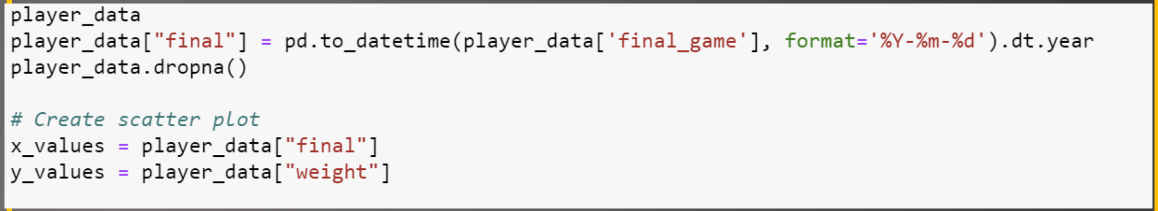
Figure 3


Ptiching ERA from 1970-2015
Earned Run Average
5.2
5.0
4.8
4.6
4.4
4.2
4.0
3.8
3.6
1970
1980
1990
2000
2010
Years

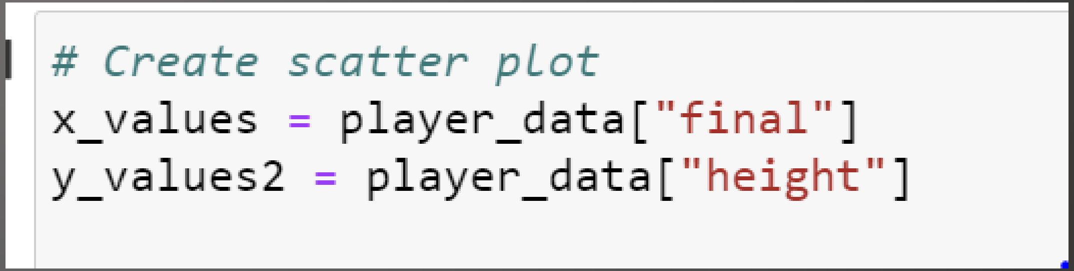
Analysis

5. What caused the era and batting average spike in the 90’s

  - Player weight and height over the years



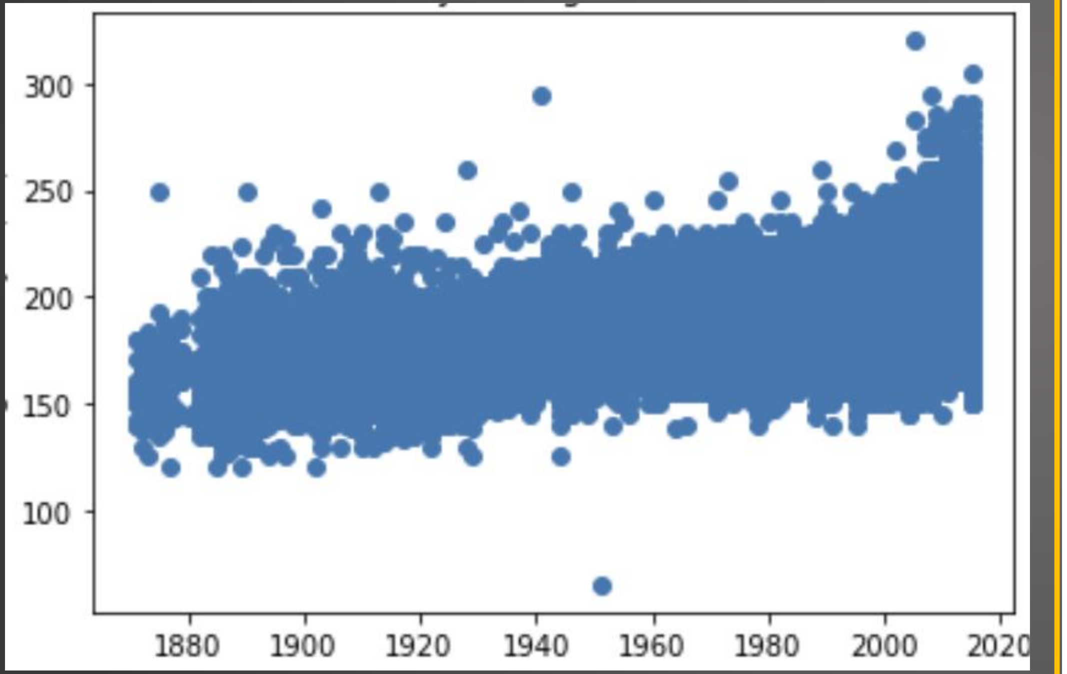
```
player_data
player_data["final"] = pd.to_datetime(player_data['final_game'], format='%Y-%m-%d').dt.year
player_data.dropna()

# Create scatter plot
x_values = player_data["final"]
y_values = player_data["weight"]
```



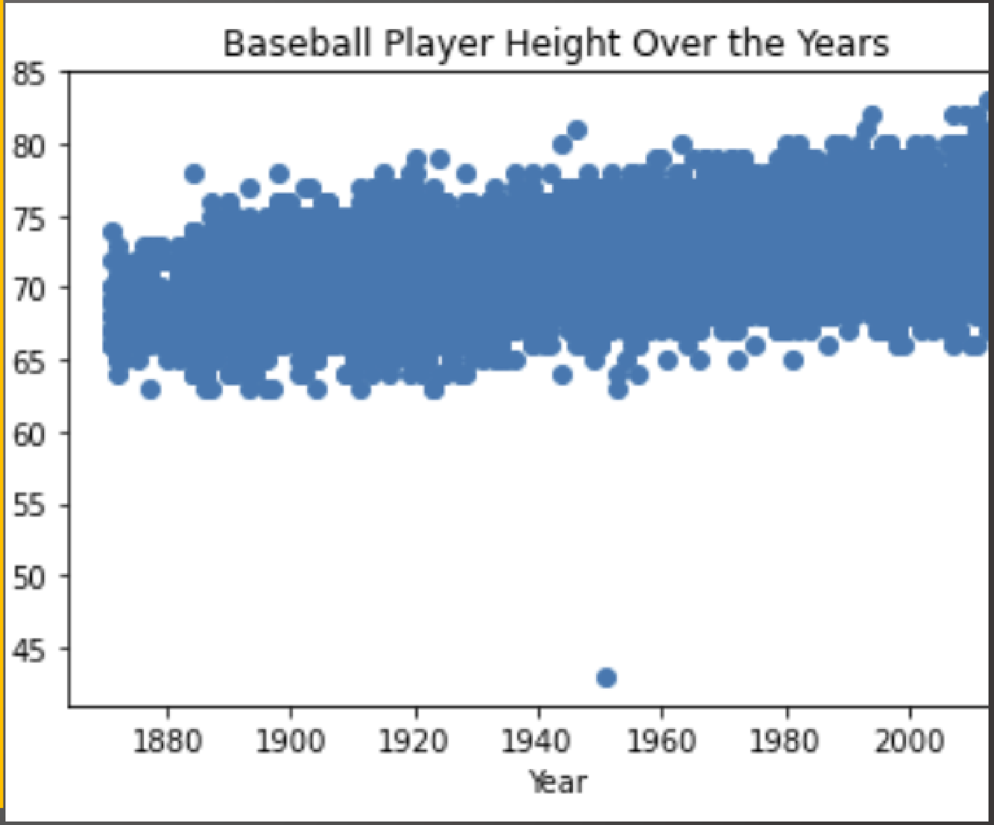
```
# Create scatter plot
x_values = player_data["final"]
y_values2 = player_data["height"]
```

Figure 4 & 5

Weight



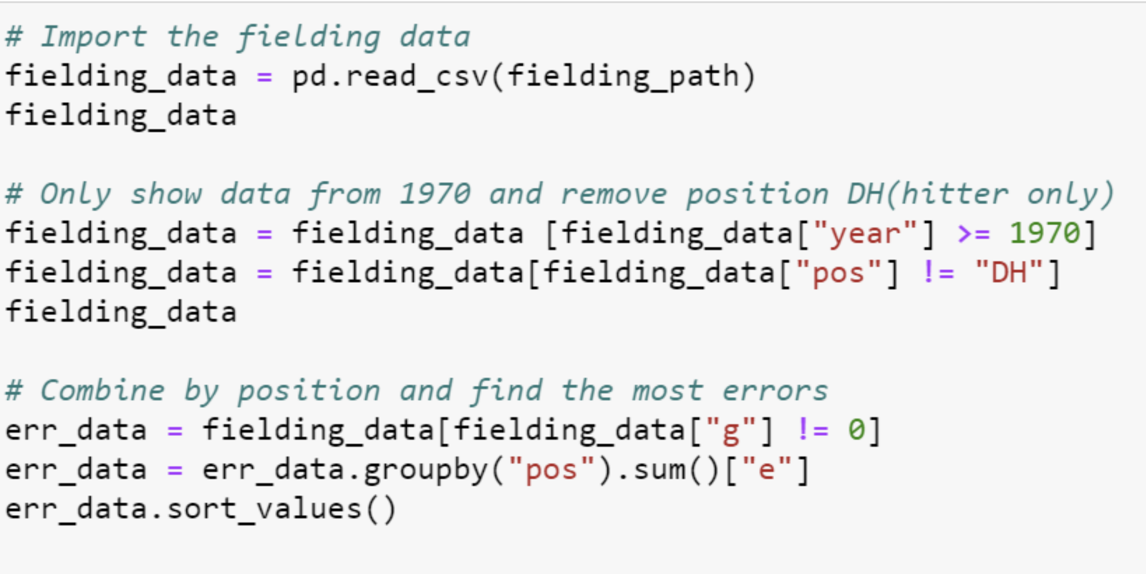
Height



Analysis

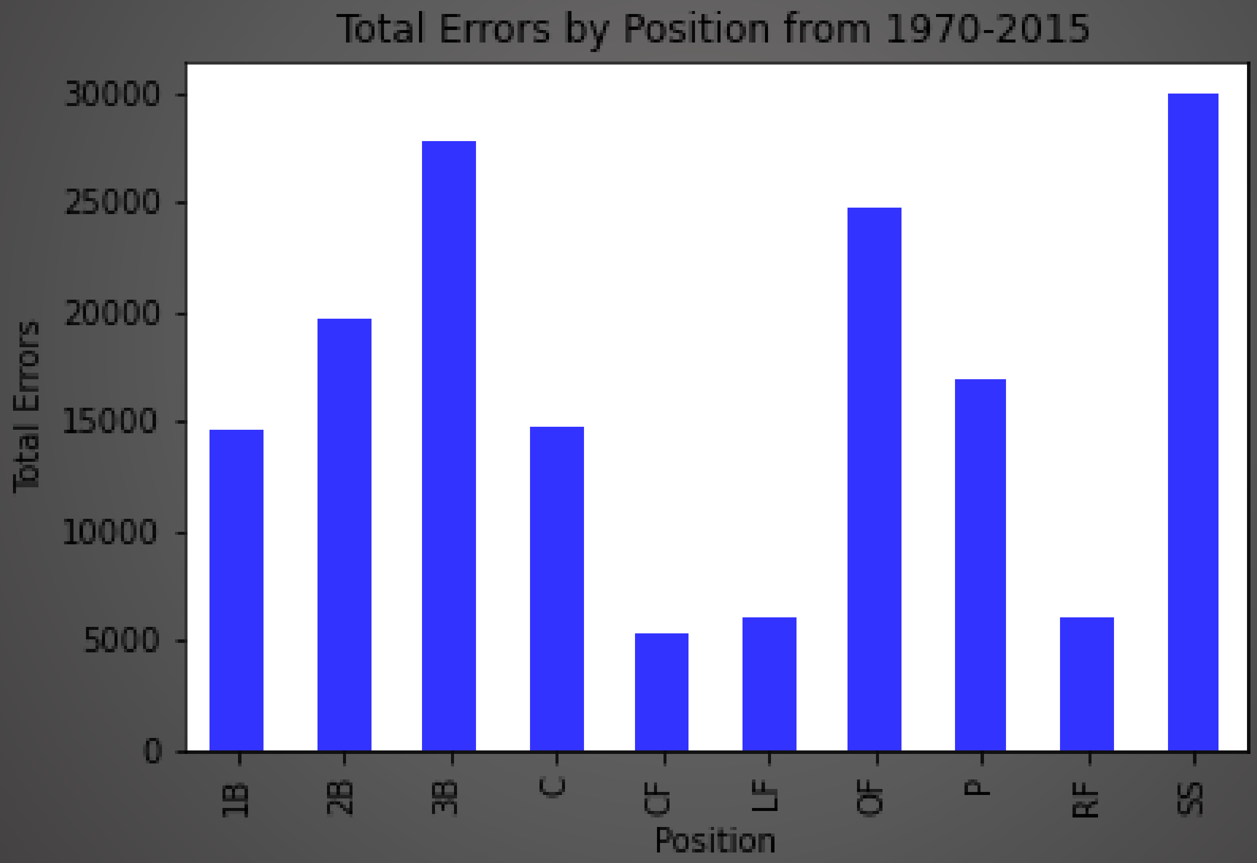
6. What position has the most errors?

- Years 1970-2015
- Removed DH position
- Sum of errors per each position



```
# Import the fielding data
fielding_data = pd.read_csv(fielding_path)
fielding_data

# Only show data from 1970 and remove position DH(hitter only)
fielding_data = fielding_data [fielding_data["year"] >= 1970]
fielding_data = fielding_data[fielding_data["pos"] != "DH"]
fielding_data

# Combine by position and find the most errors
err_data = fielding_data[fielding_data["g"] != 0]
err_data = err_data.groupby("pos").sum()["e"]
err_data.sort_values()
```

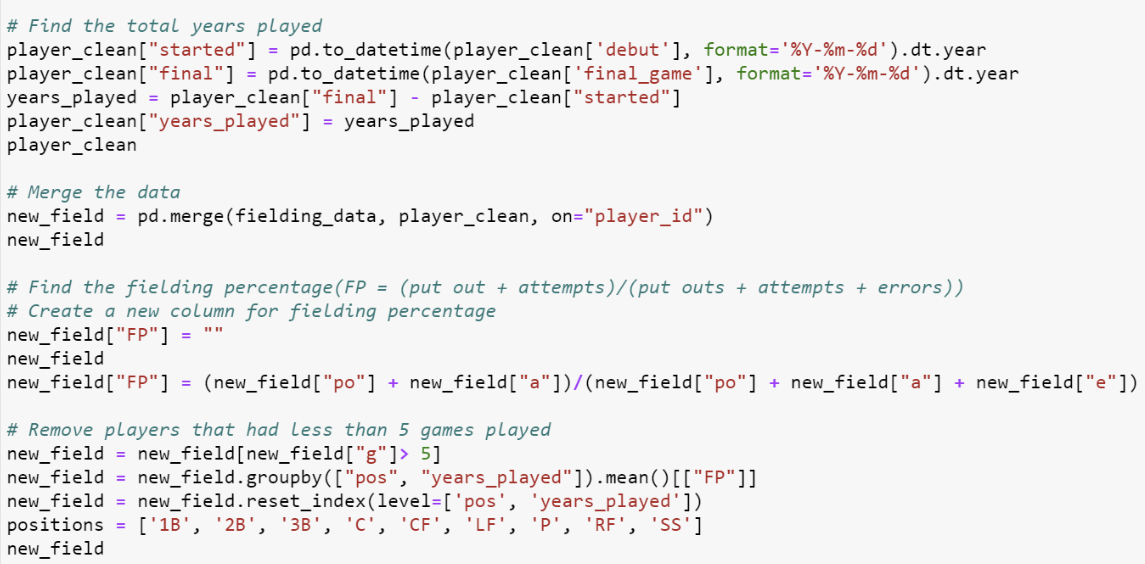
Figure 6


Total Errors by Position from 1970-2015
Total Errors
30000
25000
20000
15000
10000
5000
0
1B
2B
3B
C
CF
LF
OF
P
RF
SS
Position

Analysis

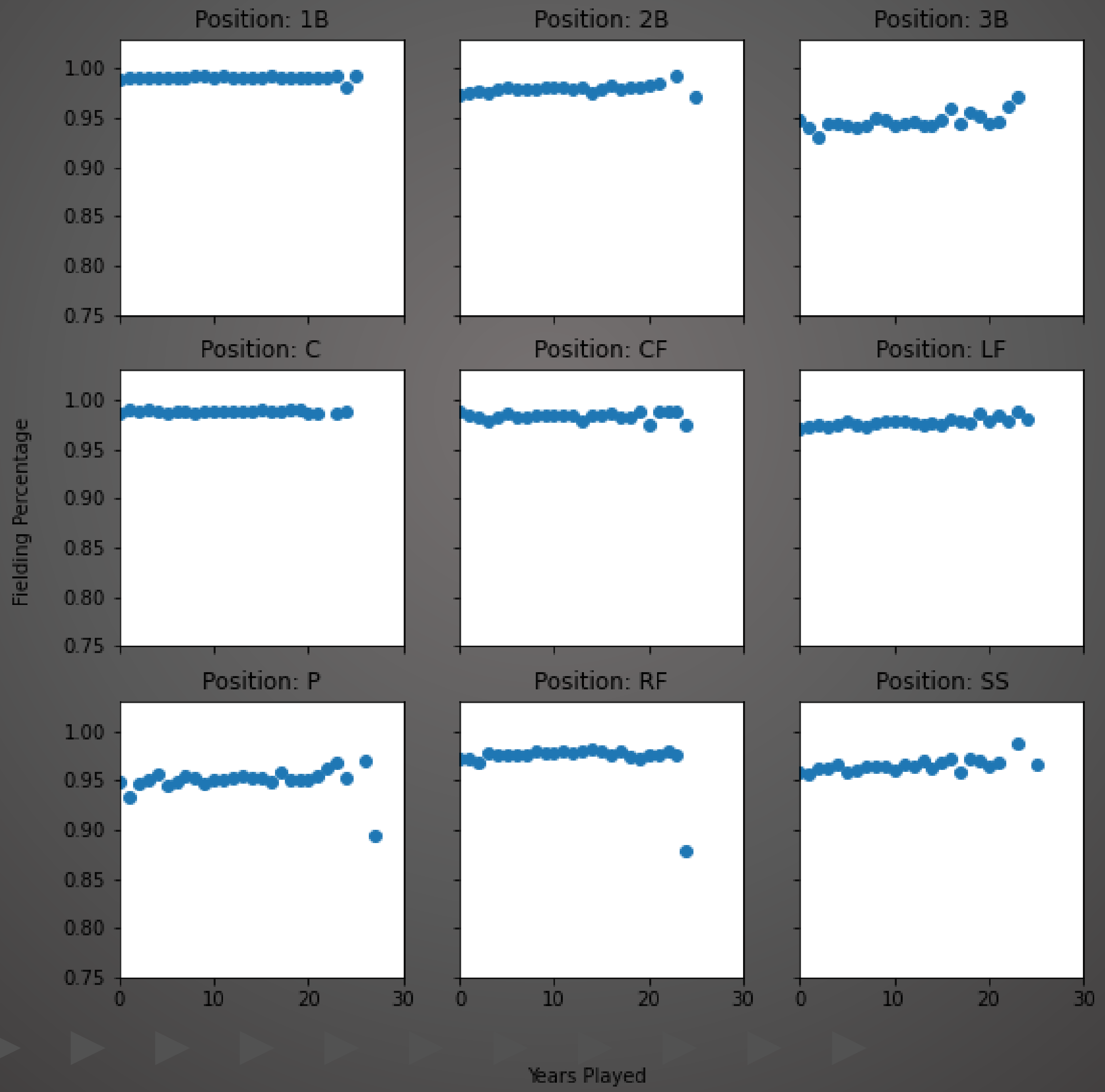
7. Does experience in professional baseball improve fielding percentage?

- Found total years played from player csv
- Merged with fielding
- Fielding percentage = (put outs + attempts)/ (put outs + attempts + errors)
- More than 5 games played



```
# Find the total years played
player_clean["started"] = pd.to_datetime(player_clean['debut'], format='%Y-%m-%d').dt.year
player_clean["final"] = pd.to_datetime(player_clean['final_game'], format='%Y-%m-%d').dt.year
years_played = player_clean["final"] - player_clean["started"]
player_clean["years_played"] = years_played
player_clean

# Merge the data
new_field = pd.merge(fielding_data, player_clean, on="player_id")
new_field

# Find the fielding percentage(FP = (put out + attempts)/(put outs + attempts + errors))
# Create a new column for fielding percentage
new_field["FP"] = ""
new_field
new_field["FP"] = (new_field["po"] + new_field["a"])/(new_field["po"] + new_field["a"] + new_field["e"])

# Remove players that had less than 5 games played
new_field = new_field[new_field["g"]> 5]
new_field = new_field.groupby(["pos", "years_played"]).mean()[["FP"]]
new_field = new_field.reset_index(level=['pos', 'years_played'])
positions = ['1B', '2B', '3B', 'C', 'CF', 'LF', 'P', 'RF', 'SS']
new_field
```

Figure 7



Position: 1B
Position: 2B
Position: 3B
Position: C
Position: CF
Position: LF
Position: P
Position: RF
Position: SS
Fielding Percentage
1.00
0.95
0.90
0.85
0.80
0.75
0
10
20
30
Years Played

Discussion

- Stats can help see how a player will measure up in the professional league
  - Is a minor league player ready for MLB
- Interesting the significant difference of where players are born
- Expected AL to have better overall batting average
- Was a trend in the 90's – 00's, but could not pinpoint to one cause
- Fielding stats very consistent
  - Small percentage make it to professional
  - If errors increase will be taken out of the game

What if?

- Would like to include post-season analysis
- Take a deeper dive into pitching stats
    - Has pitching changed over the years?
    - Are there more lefty pitchers?
    - Change in strikeouts? Or total pitches per game?
- Compare the top players over the years

Questions?