



Alfonso Herrada

QuickStart

Project 2 – Baseball Stats Analysis

MAXIMUM POINTS: 100 +(5) SECTION 1

1. Model the data to find all players who have played at least 50 games and are still active. Use the “finalGame” field from the “People” table to determine if the player is still active. What do you observe? Show steps and results.

Data was modeled in SQL by joining the ‘Peoples’ table and subquery of the ‘Appearance’ table to identify players who have played more than 50 games and are still active. ‘People_clean_active’ SQL view was created from this query.

- Appearance table subquery
 - Used window function to get an aggregated list of the sum of Total Games played for each player and then filtered for Total Games >= 50
 -
- Joined the ‘Peoples’ table to the ‘Appearance’ table subquery and filtered for ‘finalGame’ is null. Per the Lehman documentation this would indicate that the player is still active. Using SQL Case statement created a ‘Player Status’ column that would state whether a player was active/inactive based on results in ‘finalGame’. This query returned zero results.
 debut Date that player made first major league appearance
 finalGame Date that player made first major league appearance (blank if still active)
- Queried the ‘Peoples’ table alone and filtered for ‘finalGame’ is null. This produced 196 results. The results are mostly of very old players who would have played more than 80 years ago. I was able to conclude that the ‘nulls’ are null because there is no available data. The Nulls in ‘finalGame’ columns are not representative of a player's active/inactive status.

```
1 SELECT
2 playerID_people AS 'PlayerID',
3 nameFirst AS 'First Name',
4 nameLast AS 'Last Name',
5 birth_date AS 'Birth Date',
6 weight AS 'Weight',
7 height AS 'Height',
8 bats AS 'Bats',
9 throws AS 'Throws',
10 finalgame AS 'Final Game',
11 timesCompIfGyrar.birth_date,current_date) AS 'Current Age'
12 FROM people
13 WHERE finalgame_date is null
14
```

PlayerID	First Name	Last Name	Birth Date	Weight	Height	Bats	Throws	Final Game	Current Age
acton99	Manny	Acta	1989-01-11	172	74	R	R	00000	35
adair99	Bill	Adair	1915-02-10	172	74	R	R	00000	109
arnold99	Bill	Arnold	1909-09-03	00000	00000	00000	00000	00000	113
barber99	Frank	Barcroft	1946-01-09	00000	00000	00000	00000	00000	78
barke99	Al	Barck	1914-04-02	185	71	00000	00000	00000	107
barre99	Ed	Barrow	1904-06-10	00000	00000	00000	00000	00000	119
bell99	Cool Papa	Bel	1903-05-17	155	72	R	L	00000	118
benet99	Terry	Berington	1946-07-07	00000	00000	R	R	00000	65
benesh99	Hugo	Benesh	1989-04-01	00000	00000	00000	00000	00000	35
boland99	John	Boland	1948-08-19	00000	00000	00000	00000	00000	75
bradley99	Dave	Bradley	1923-06-23	175	71	R	R	00000	101
browns99	Daren	Brown	1967-06-13	185	76	R	R	00000	57
brown99	Freeman	Brown	1941-01-31	00000	00000	00000	00000	00000	83
brown99	Ray	Brown	1908-03-23	175	71	R	R	00000	116
buckley99	Al	Buckley	1961-01-11	00000	00000	00000	00000	00000	63
bullock99	Morgan	Bullock	1927-12-26	00000	00000	00000	00000	00000	96
burns99	Walt	Burns	1965-01-10	00000	00000	00000	00000	00000	59
butler99	Clarence	Butler	1924-11-18	00000	00000	00000	00000	00000	99
byrnes99	Charlie	Byrnes	00000	00000	00000	00000	00000	00000	00000
cantlon99	Joe	Cantlon	1961-01-19	00000	00000	00000	00000	00000	63
carlson99	Alexander	Carlson	1920-01-19	00000	00000	00000	00000	00000	103
caylo99	Ole	Caylor	1949-12-14	00000	00000	00000	00000	00000	75

Result 9
104 12:19:00 SELECT playerID_people AS 'PlayerID', nameFirst AS 'First Name', nameLast AS 'Last Name', birth_date AS 'Birth Date', weight ... 196 row(s) returned



Alfonso Herrada
QuickStart
Project 2 – Baseball Stats Analysis

- The 'finalGame' column actually shows the date of the respective player's last appearance in a game. This column should be more appropriately named, lastGamePlayed.
- Therefore, in order to determine if a player is active or inactive. I used the beginning of the 2019 baseball season, 03/28/2019, as my line of demarcation.
 - Active players = finalGame >= 2019-03-28
 - Inactive players = finalGame < 2019-03-28
 - Using this logic we assume that a player who did not play in the 2019 season is considered inactive. A player could have been injured and not played the entire 2019 season but returned healthy for the 2020 season.
- This gave me a total of 1001 active players who played at least 50 games in their careers.

2. Model the data to find all players who have played at least 50 games and are inactive. Retrieve weights, throws, bats, all birth-related and name related columns from the "People" table and retrieve all columns from the batting table.

Data was modeled in SQL. I created a SQL view, 'People_clean_all', this is similar to 'people_clean_active' except that, 'People_clean_all', does not filter by 'finalGame' so active and inactive players are both included. This view was then filtered for Player_Status = Inactive

PlayerID	First Name	Last Name	Birth Date	Weight	Height	Bats	Throws	Total_Games	Final Game	Current Age	Player_Status
suzukic01	Ichiro	Suzuki	1973-10-22	175	71	L	R	2653	2019-03-21	48	Inactive
hollima01	Matt	Holiday	1980-01-15	240	76	R	R	1903	2018-10-01	42	Inactive
knebeco01	Corey	Knebel	1991-11-26	220	76	R	R	224	2018-10-01	30	Inactive
garcoja02	Jaime	Garcia	1986-07-08	215	74	L	L	222	2018-10-01	35	Inactive
beltrad01	Adrian	Beltre	1979-04-07	220	71	R	R	2933	2018-09-30	43	Inactive
utleych01	Chase	Utley	1978-12-17	195	73	L	R	1937	2018-09-30	43	Inactive
phillbr01	Brandon	Phillips	1981-06-28	211	72	R	R	1902	2018-09-30	40	Inactive
reyesjo01	Jose	Reyes	1983-06-11	195	72	B	R	1877	2018-09-30	38	Inactive
mauerjo01	Joe	Mauer	1983-04-19	225	77	L	R	1858	2018-09-30	39	Inactive
bautijo02	Jose	Bautista	1980-10-19	205	72	R	R	1798	2018-09-30	41	Inactive
escobal02	Alcides	Escobar	1986-12-16	205	73	R	R	1437	2018-09-30	35	Inactive
spande01	Denard	Span	1984-02-27	210	72	L	L	1359	2018-09-30	38	Inactive
jacksau01	Austin	Jackson	1987-02-01	198	73	R	R	1115	2018-09-30	35	Inactive
blancgr01	Gregor	Blanco	1983-12-24	187	70	L	L	1060	2018-09-30	38	Inactive
perezsa02	Salvador	Perez	1990-05-10	240	76	R	R	942	2018-09-30	31	Inactive
gattiev01	Evan	Gattis	1986-08-18	270	76	R	R	706	2018-09-30	35	Inactive
belisma01	Matt	Belisle	1980-06-06	230	75	R	R	694	2018-09-30	41	Inactive
ellisaj01	A. J.	Ellis	1981-04-09	225	74	R	R	672	2018-09-30	41	Inactive
rosalad01	Adam	Rosales	1983-05-20	200	74	R	R	653	2018-09-30	38	Inactive
rosalad02	Adam	Rosales	1983-05-20	200	74	R	R	653	2018-09-30	38	Inactive

Result 3 x

Output

Action Output

#	Time	Action	Message
27	13:05:03	SELECT * -- Peoples and appearance tables joined. Games played >= 50 and Player is active. FROM (SELECT -- Peoples and Appearance...	10278 row(s) returned



Alfonso Herrada

QuickStart

Project 2 – Baseball Stats Analysis

3. From (2), add a calculated column with the players age and a calculated column with each player's first and last name concatenated.

Joined 'Batting' table with 'People_clean_all' SQL view. Created calculated columns in 'People_clean_all'

- Players Age -
 - CASE WHEN death_date IS NULL THEN timestampdiff(year, p.birth_date, CURRENT_DATE) ELSE 'Passed_Away' END AS 'Current_Age'
- First and Last Name-
 - concat(nameFirst, " ", nameLast) AS 'Name'

4. Once the calculated columns are added, drop the other columns related to birth date and name.

Created SQL view, 'batting_people_clean'

```
32 FROM batting b
33 Join people_clean_all p
34 on b.playerID_batting = p.playerID
35
```

playerID_batting	Name	Current_Age	weight	height	bats	throws	Total_Games	finalgame	Player_Status	teamID_batting	Year	GA	AB	BA	RBI	HR	1B	2B	3B	Stolen Bases	Caught Stealing	Base on Balls	Strikeouts	Ints Wa
aardsda01	David Aardema	40	215	75	R	R	331	2015-08-23	Inactive	SPN	2004	11	0	0.000	0	0	0	0	0	0	0	0	0	0
aardsda01	David Aardema	40	215	75	R	R	331	2015-08-23	Inactive	CHN	2006	45	2	0.000	0	0	0	0	0	0	0	0	0	0
aardsda01	David Aardema	40	215	75	R	R	331	2015-08-23	Inactive	CHA	2007	25	0	0.000	0	0	0	0	0	0	0	0	0	0
aardsda01	David Aardema	40	215	75	R	R	331	2015-08-23	Inactive	BOS	2008	47	1	0.000	0	0	0	0	0	0	0	0	1	0
aardsda01	David Aardema	40	215	75	R	R	331	2015-08-23	Inactive	SEA	2009	73	0	0.000	0	0	0	0	0	0	0	0	0	0
aardsda01	David Aardema	40	215	75	R	R	331	2015-08-23	Inactive	SEA	2010	53	0	0.000	0	0	0	0	0	0	0	0	0	0
aardsda01	David Aardema	40	215	75	R	R	331	2015-08-23	Inactive	NYA	2012	1	0	0.000	0	0	0	0	0	0	0	0	0	0
aardsda01	David Aardema	40	215	75	R	R	331	2015-08-23	Inactive	NYN	2013	43	0	0.000	0	0	0	0	0	0	0	0	0	0
aardsda01	David Aardema	40	215	75	R	R	331	2015-08-23	Inactive	ATL	2015	33	1	0.000	0	0	0	0	0	0	0	0	1	0
aaronha01	Hank Aaron	88	180	72	R	R	3298	1976-10-03	Inactive	ML1	1954	122	468	0.280	69	13	85	27	6	2	2	28	39	N/A
aaronha01	Hank Aaron	88	180	72	R	R	3298	1976-10-03	Inactive	ML1	1955	153	602	0.314	106	27	116	37	9	3	1	49	61	5
aaronha01	Hank Aaron	88	180	72	R	R	3298	1976-10-03	Inactive	ML1	1956	153	609	0.328	92	26	126	34	14	2	4	37	54	6
aaronha01	Hank Aaron	88	180	72	R	R	3298	1976-10-03	Inactive	ML1	1957	151	615	0.322	132	44	121	27	6	1	1	57	58	15
aaronha01	Hank Aaron	88	180	72	R	R	3298	1976-10-03	Inactive	ML1	1958	153	601	0.326	95	30	128	34	4	4	1	59	49	16
aaronha01	Hank Aaron	88	180	72	R	R	3298	1976-10-03	Inactive	ML1	1959	154	629	0.355	123	39	131	46	7	8	0	51	54	17
aaronha01	Hank Aaron	88	180	72	R	R	3298	1976-10-03	Inactive	ML1	1960	153	590	0.292	126	40	101	20	11	16	7	60	63	13
aaronha01	Hank Aaron	88	180	72	R	R	3298	1976-10-03	Inactive	ML1	1961	155	603	0.327	120	34	114	39	10	21	9	56	64	20
aaronha01	Hank Aaron	88	180	72	R	R	3298	1976-10-03	Inactive	ML1	1962	156	592	0.323	128	45	112	28	6	15	7	66	73	14
aaronha01	Hank Aaron	88	180	72	R	R	3298	1976-10-03	Inactive	ML1	1963	161	631	0.319	130	44	124	29	4	31	5	78	94	18
aaronha01	Hank Aaron	88	180	72	R	R	3298	1976-10-03	Inactive	ML1	1964	145	570	0.328	95	24	131	30	2	22	4	62	46	9
aaronha01	Hank Aaron	88	180	72	R	R	3298	1976-10-03	Inactive	ML1	1965	150	570	0.318	89	32	108	40	1	24	4	60	81	10
aaronha01	Hank Aaron	88	180	72	R	R	3298	1976-10-03	Inactive	ATL	1966	158	603	0.279	127	44	100	23	1	21	3	76	96	15

SECTION 2

Answer the following questions:

1. Which active player had the most runs batted in (RBI from the batting table) from 2015 to 2018?

Nolan Arenado 503 RBI's



Alfonso Herrada QuickStart Project 2 – Baseball Stats Analysis

people_clean_active* batting_clean* RBI_Leader_2015_2018

```
1 • Select -- Active Players Total_RBI from 2015 - 2018 desc. Joining of people_clean_active and batting_clean
2 distinct playerID_people,
3 nameFirst,
4 nameLast,
5 SUM(RBI) OVER (PARTITION BY playerID_batting) AS Total_RBI
6 FROM
7 (
8   SELECT
9     * -- Peoples and appearance tables joined. Games played >= 50 and Player is active.
10    FROM
11    (
12      SELECT
13        -- Peoples and Appearance tables joined. Games played >= 50 and
14        p.playerID_people,
```

Result Grid

playerID_people	nameFirst	nameLast	Total_RBI
arenano01	Nolan	Arenado	503
donaipo02	Josh	Donaldson	323
davisch02	Chris	Davis	311
encarned01	Edwin	Encarnacion	452
goldspa01	Paul	Goldschmidt	408
moralle01	Kendrys	Morales	341
martijo02	J. D.	Martinez	404
abreujo02	Jose	Abreu	381
rizzoan01	Anthony	Rizzo	420
kempma01	Matt	Kemp	357
bravabr01	Kris	Brvant	326

Result 25

2. How many double plays did Albert Pujols ground into (“GIDP” from Batting table) in 2016?

24

people_clean_active* batting_clean* Pujols_GIDP_2016*

```
1 • Select -- Albert Pujols 2016 Ground Into Double Play stats. Joining of people_clean_active and batting_clean
2 nameFirst,
3 nameLast,
4 Year,
5 GIDP
6 FROM
7 (
8   SELECT
9     * -- Peoples and appearance tables joined. Games played >= 50 and Player is active.
10    FROM
11    (
12      SELECT
13        -- Peoples and Appearance tables joined. Games played >= 50 and
14        p.playerID_people,
```

Result Grid

nameFirst	nameLast	Year	GIDP
Albert	Pujols	2016	24

3. In which year were the highest number of Hall of Fame awards given?

In 2006 there were 18 people inducted into the Hall of Fame.



Alfonso Herrada
QuickStart
Project 2 – Baseball Stats Analysis

```
1 • SELECT
2 distinct yearid,
3 count(inducted) over (partition by yearid order by yearid) AS HOF
4 FROM lahmansbaseballdb.halloffame
5 where inducted = 'y'
6 order by HOF desc
```

<		
Result Grid		
Filter Rows: <input type="text"/>		
Export:		
Wrap Cell Content:		
yearid	HOF	
2006	18	
1946	11	
1939	10	
1945	10	
1937	8	

4. In what category were the highest number of Hall of Fame awards given?

The Players category with 256 HOF inductees.

```
1 • SELECT
2 distinct category,
3 count(category) over (partition by category) AS HOF
4 FROM lahmansbaseballdb.halloffame
5 where inducted = 'y'
6 order by HOF desc
```

<		
Result Grid		
Filter Rows: <input type="text"/>		
Export:		
Wrap Cell Center		
category	HOF	
Player	256	
Pioneer/Executive	34	
Manager	23	
Umpire	10	

5. In which year were the highest number of Wild Pitches recorded?

In 1884 there were 2,183 wild pitches.



Alfonso Herrada
QuickStart
Project 2 – Baseball Stats Analysis

```
1 • SELECT
2   distinct yearID,
3   sum(WP) over (partition by yearID ) as WP_Year_Total
4 FROM lahmansbaseballdb.pitching
5 order by WP_Year_Total desc
6
```

Result Grid | Filter Rows: | Export: | Wrap Cell Conte

	yearID	WP_Year_Total
▶	1884	2183
	2018	1847
	2017	1810
	2016	1808
	2019	1788
	2015	1758
	2013	1736
	2014	1696
	2010	1674
	1999	1632
	1998	1603
	2009	1600
	1890	1598
	2008	1576
	----	----

6. Name the player inducted in the hall of fame with the highest number of industry experience?

Pitchers do not pitch every game like players who play other positions would/could. Depending on the pitcher's role they generally appear in every 3-5 games. Therefore, I analyzed pitchers and non pitchers in separate groups.

Non- Pitchers

Max Experience - Carl Yastrzemski appeared in 3304 games

Min. Experience - Monte Irvin appeared in 764 games

Pitchers

Max Experience - Dennis Eckersley appeared in 1071 games

Min. Experience - Satchel Paige appeared in 179 games

SECTION 3

Create the following plots:

- A histogram of triples(3B) per year. What do you observe?



Alfonso Herrada
QuickStart
Project 2 – Baseball Stats Analysis

- **Create a scatter plot relating triples (3B) and steals (SB). What do you observe? Now calculate the correlation between these two variables.**

Used Power BI to plot and complete correlation analysis. Correlation coefficient was calculated using the Quick Measure feature. There is a low to mid correlation between 3B and SB overall. The dashboard also has a card illustrating the coefficient which updates as the dashboard is filtered.

SECTION 4 (Bonus) (5 Points)

Show four DAX Calculations in your data model. You can create measures of your choice as long as they relate to the data and provide valuable insights.

